Hydrological Processes

A multi-objective approach to select hydrological models and constrain structural uncertainties for climate impact assessments

Journal:	*Hydrological Processes*
Manuscript ID	Draft
Wiley - Manuscript type:	Special Issue Paper
Date Submitted by the Author:	n/a

Complete List of Authors:	Saavedra, Danny; Universidad de Chile, Department of Civil Engineering Mendoza, Pablo; Universidad de Chile, Department of Civil Engineering; Universidad de Chile, Advanced Mining Technology Center Addor, Nans; University of Exeter, Geography Llauca, Harold; Servicio Nacional de Meteorología de Hidrología del Perú, Hydrology Vargas, Ximena; Universidad de Chile, Department of Civil Engineering

Keywords:	hydrologic change, model structure uncertainty, modular modelling framework, Pareto scheme, signatures, hydrological consistency

A multi-objective approach to select hydrological models and constrain structural uncertainties for climate impact assessments

Danny Saavedra[1], Pablo A. Mendoza[1,2], Nans Addor[3], Harold Llauca[4], and Ximena Vargas[1]

[1] Department of Civil Engineering, Faculty of Physical and Mathematical Sciences, Universidad de Chile, Chile.

[2] Advanced Mining Technology Center (AMTC), Faculty of Physical and Mathematical Sciences, Universidad de Chile, Chile.

[3] Geography, College of Life and Environmental Sciences, University of Exeter, UK.

[4] Servicio Nacional de Meteorología e Hidrología del Perú (SENAMHI), Lima, Perú.

Corresponding Author: Pablo A. Mendoza (pamendoz@uchile.cl)

Abstract

The assessment of climate change impacts on water resources and flood risk is typically underpinned by hydrological models calibrated and selected based on observed streamflow records. Yet, changes in climate are rarely accounted for when selecting hydrological models, which compromises their ability to robustly represent future changes in catchment hydrology. In this paper, we test a simple framework for selecting an ensemble of calibrated hydrological model structures in catchments where changing climatic conditions have been observed. We start by considering 78 model structures produced using the FUSE modular modelling framework and rely on a Pareto scheme to select model structures maximizing model efficiency in both wet and dry periods. The application of this approach in three case study basins in Peru enables the identification of structures with good robustness, but also good performance according to hydrological signatures not used for model selection. We also highlight that some model structures that perform well according to traditional efficiency metrics have low performance in contrasting climates or suspicious internal states and fluxes. Importantly, the model selection approach followed here helps to reduce the spread in precipitation elasticities and temperature sensitivities, providing a clearer picture of future hydrological changes. Overall, this work demonstrates the potential of using contrasting climatic conditions in a multi-objective framework to produce robust and credible simulations, and to constrain structural uncertainties in hydrological projections.

Keywords: Hydrologic change; model structure uncertainty; modular modelling framework; Pareto scheme; signatures; hydrological consistency.

1. INTRODUCTION

Climate change is greatly affecting the economy and quality of life of populations around the world, causing an increase in the frequency and intensity of extreme hydrological events such as droughts and floods (Gavrilović et al., 2012; Correa et al., 2017; Shiru et al., 2019; Haile et al., 2019; Son et al., 2020; Bhardwaj et al., 2020; Alvarez-Garreton et al., 2021), challenging thus water resources management and flood risk. The hydrological community is tackling this challenge and, in a collaborative effort, continuous improvements are being developed in the methodologies used in the assessment of climate change impacts.

The assessment of climate change impacts on water resources commonly involves several methodological choices, which include the selection of emission scenarios, global climate models (GCM), initial conditions, downscaling method, hydrologic model structure and parameter values (e.g., Wilby & Harris, 2006; Chen et al., 2011; Addor et al., 2014; Chegwidden et al., 2019). The above decisions lead to uncertainties, whose relative importance may differ depending on specific hydroclimatic conditions and basin characteristics (Clark et al., 2016). In particular, many authors have found that the choice of the hydrologic model structure (i.e., choice of processes explicitly represented, model parameterizations, architecture and connectivity) and the choice of parameters (i.e., the coefficients in model equations, either free or observable) may have large effects on the characterization of climate change impacts (e.g., Miller et al., 2012; Seiller et al., 2012; Vano et al., 2012; Brigode et al., 2013; Mendoza et al., 2015, 2016; Fowler et al., 2018a; Melsen et al., 2018).

The sole effects of hydrologic model choice – commonly based on legacy, rather than adequacy (Addor & Melsen, 2019) – have been widely explored in the context of climate change impacts. A large body of work has relied on the selection of a small ensemble of hydrological models (e.g., Jiang et al., 2007; Bae et al., 2011; Mendoza et al., 2015; Mizukami et al., 2016), while a few authors have proposed explicit changes in model structures to address this issue (Westra et al., 2014; Grigg & Hughes, 2018). The assessment of structural uncertainties is now facilitated by the emergence of modular modelling frameworks (MMFs; e.g., Pomeroy et al., 2007; Clark et al., 2008, 2015; Kavetski & Fenicia, 2011; Niu et al., 2011; Coxon et al., 2019; Knoben et al., 2019; Craig et al., 2020), which allow to design controlled experiments to test hypotheses about catchment functioning (Clark et al., 2011). Of course, the large number of modelling options available in MMFs raises the challenge to sample model space in order to efficiently capture structural uncertainty (Remmers et al., 2020), especially under scenarios of changing climatic conditions.

During decades, the hydrology community adopted differential split-sample tests (DSST; Klemes 1986) as a standard practice to assess the temporal stability of model performance. However, many authors have reported decreased skill when a model is applied in very different climatic conditions compared to those used to infer the parameter values (e.g., Vaze et al., 2010; Merz et al., 2011; Seiller et al., 2012; Brigode et al., 2013; Motavita et al., 2019; Pan et al., 2019). For example, Coron et al. (2012) examined the extrapolation capacity of three hydrological models for different climatic conditions in 216 basins in Australia, proposing a generalized split-sample test (GSST) based on the DSST methodology. Their results demonstrated that the transfer of model parameters in time can introduce errors in simulations, and therefore lack of robustness when the models are used in a changing climate. Stephens et al. (2019) developed three experiments to test the potential of (1) transferring model parameters in time, (2) improve simulations under future climate scenarios, and (3) varying model parameters according to climate conditions for improved simulations.

Stephens et al. (2019) used the GR4J (Perrin et al., 2003) conceptual model to conduct three experiments over 164 Australian basins, obtaining mixed results in their experiments to improve model performance under contrasting climatic conditions. More recently, Duethmann et al. (2020) analysed the causes of the low performance of a semi-distributed hydrological model under changing climate conditions over a large number of basins in Australia, where they mainly focused on: (1) data problems, (2) problems related to the model calibration and (3) model structure deficiencies. Duethmann et al. (2020) found that poor model performance is mainly because most model structures ignore changes in vegetation dynamics, and due to temporal inhomogeneities in precipitation data.

Fowler et al. (2016) revisited the problem of temporal instability in model performance through the application of a Pareto framework – aimed to find parameter sets that simultaneously maximize model efficiency in a wet and a dry period – in 85 catchments located in Australia, using five conceptual hydrological models. They found that temporal instabilities reported in previous studies may be attributed to poor parameter estimation strategies, rather than model structural inadequacies. Fowler et al. (2018a) assessed hydrological model performance in historical multi-year droughts, finding that they often have poor performance that could be attributed to errors in the data, model structure or parameter values. It should be noted that their Pareto scheme was used to discriminate parameter sets given a fixed model structure, rather than screening competing models with the same application purpose.

Hydrologic sensitivities to climate perturbations are attractive due to their simple formulation, and because they allow quick estimates of runoff production under different climate scenarios (Vano & Lettenmaier, 2014; Vano et al., 2015; Milly et al., 2018; Lehner et al., 2019). Vano et al. (2012) examined the effects of hydrologic model choice on the estimation of precipitation elasticities and temperature sensitivities in the Colorado River basin, finding large differences attributed to structural discrepancies and model biases, since the models were not configured to simulate streamflow. Mendoza et al. (2015) compared inter-model differences in projected hydrologic changes before and after conducting parameter estimation, concluding that large differences remain between calibrated models. Nevertheless, the subjective choice of models in these studies, the lack of further model evaluation in contrasting climatic periods, and the body of work previously referred to reinforces the urgency to (1) improve parameter estimation and model selection strategies, and (2) conduct plausibility checks in model structures to obtain coherent results under changing climatic conditions. In this paper, we combine elements from recent studies to address the following questions:

- Can the hydrological consistency of simulations in contrasting climatic periods be improved by sampling the model space with a simple Pareto framework?
- Can links be drawn between the components of the models selected by this procedure?
- Can this model selection procedure reduce uncertainties in precipitation elasticities and temperature sensitivities?

Hence, we propose an approach based on (1) selecting dry and wet sub-periods, (2) calibrating hydrological models in each sub-period, (3) choosing the combinations of hydrological model and parameter set that maximize performance in wet and dry years, (4) assessing hydrological consistency and screening models, and (5) quantifying the ensemble spread in hydrologic sensitivities resulting from the model subset. We apply this framework in three basins located in Peru and rely on model structures produced using the Framework for Understanding Structural Errors (FUSE; Clark et al., 2008).

2. STUDY DOMAIN AND DATA

The National Water Authority (ANA, by its acronym in Spanish) of Peru has identified three main hydrographic regions (ANA 2009) – the Pacific region, the Atlantic or Amazon region and Lake Titicaca region – with different geomorphological, climatic and hydrological characteristics (see Lavado Casimiro et al., 2012; Heidinger et al., 2018; Rau et al., 2018). To represent such diversity, we select three case study basins (Figure 1): two of these are located in Peru (Vilcanota and Huancane), and the other is a transboundary catchment located between Peru and Ecuador (Puyango-Tumbes). The Vilcanota River basin belongs to the Atlantic region, the Huancane River basin belongs to the Lake Titicaca region, and the Puyango-Tumbes River basin belongs to the Pacific region. During recent years, these basins have seen an increase in flooding during the rainy season, and water shortages during the dry season (Sanabria et al. 2009; Lavado Casimiro et al. 2011; Rivas & Rivas, 2013; Andres et al. 2014; Zulkafli, 2014; Takahashi & Martínez Grimaldo, 2015; Martínez & Céspedes, 2017).

[Insert Figure 1]

Table 1 summarizes the main physiographic and hydroclimatic characteristics of the three basins over the period Sep/1986 – Aug/2016 (note that the water year in Peru starts in September and ends in August). Annual runoff ratios (RR) and the mean annual runoff in the Puyango-Tumbes and Huancane basins were estimated over shorter periods, due to the lack of streamflow records in some years. The Vilcanota and Huancane River basins (located in the southeast) have similar mean elevations, while the Puyango-Tumbes River basin (located in the northwest, close to the equatorial line) has the lowest mean elevation (Figure 1), and higher values of mean annual precipitation, aridity index and runoff ratio (0.98), which implies that a large fraction of precipitation contributes to runoff. On the other hand, the Huancane River basin has the lowest runoff ratio (0.25), which implies a high evaporative fraction. The Puyango-Tumbes River basin has a high aridity index in comparison with the other catchments, since the basin has a great potential to evaporate.

[Insert Table 1]

Many authors have argued that the lack of meteorological observations in the Peruvian Andes and the Amazonia is the main limitation for climate change impact studies (Lavado Casimiro et al., 2011; Andres et al., 2014; Zulkafli, 2014; Vegas Galdos et al., 2015; Aybar et al., 2019; Rau et al., 2018; Llauca et al., 2021). To address this issue, Aybar et al. (2019) and Huerta et al. (2018) developed the Peruvian Interpolated data of SENAMHI's Climatological and hydrological Observations (PISCO) database, which provides daily time series of precipitation, minimum and maximum temperature for the period 1981-2016, with a 0.1° horizontal resolution. Precipitation time series in PISCOp (Aybar et al., 2019) were obtained using geostatistical and deterministic interpolation methods that include three precipitation sources: (1) the quality-controlled and infilled national rain gauge dataset, (2) radar-gauge merged precipitation climatologies, and (3) the Climate Hazards Group Infrared Precipitation (CHIRP) estimates. Daily time series of maximum and minimum temperature in PISCOt (Huerta et al., 2018) were obtained from: (1) observed maximum and minimum temperature data, (2) soil temperature product from the MODIS sensor (Moderate Resolution Imaging Spectroradiometer), and (3) geographic predictors (e.g., elevation, longitude, latitude and Topographic Dissection Index). The reader is referred to Huerta et al. (2018) for full descriptions on the development of temperature products. The PISCO product is freely available through the IRI Data Library website (http://iridl.ldeo.columbia.edu/SOURCES/.SENAMHI/.HSR/.PISCO/).

Daily streamflow time series were obtained from the Peruvian National Meteorological and Hydrological Service (Servicio Nacional de Meteorología e Hidrología, SENAMHI). For the Vilcanota River basin, the Intihuatana Km-105 station provides streamflow data for the period 1985 – 2016; for the Puyango-Tumbes River basin, we collect data from the El Tigre station for the period 1981 - 2016, and the Huancane bridge station provides streamflow records for the period 1988 - 2016.

3. APROACH

The proposed method to select model structures under changing climatic conditions is outlined in Figure 2, and includes the following steps: (1) selection of wet and dry periods, (2) calibration of all model structures in each basin, (3) sampling the model space based on temporally consistent performance skill, (4) assessment of hydrological consistency and model screening, and (5) quantifying the effects of model sampling on the spread in precipitation elasticities and temperature sensitivities. The step (3) builds upon the approach proposed by Fowler et al. (2018a), which aims to maximize model performance in both wet and dry periods using a Pareto analysis scheme. Hence, the framework involves the selection of two 5-year periods with contrasting climates (section 3.1.1), and a one-year period for model spin-up. Once these subperiods are defined, we set up a multi-model (MM) ensemble that we sequentially refine and transform into a smaller ensemble that provides hydrologically consistent simulations through the following steps:

i. MM0-dry and MM0-wet: full FUSE model ensemble (i.e., 78 model structures) calibrated in dry and wet periods by minimizing RMSE (section 3.1.2).
ii. MMP-dry and MMP-wet: a small 5-member ensemble obtained after applying the Pareto scheme framework with the evaluation metrics listed in Table 4 (section 3.1.3).
iii. MMPS-dry and MMPS-wet: Same as 2, but after screening MMP-dry and MMP-wet based on the seasonal behaviour of internal states and fluxes (section 3.1.4).

The details of each step are described in the following sub-sections.

3.1 Selection of analysis periods

Based on the differential split-sample test (DSST) procedure formulated by Klemes (1986), we select wet and dry analysis periods that include five consecutive years, using runoff and mean annual precipitation time series (see Figure 2, box 1). To verify the contrast between these periods, we examine annual hydroclimatic characteristics (e.g., mean annual runoff, mean annual precipitation, mean annual PET, runoff ratio and aridity index), seasonal variations in some fluxes, and the daily flow duration curve. The selected periods are displayed in Figure 3. For the Vilcanota River basin (Figure 3, left panels), we consider Sep/1988-Aug/1993 as dry period, and Sep/1999 - Aug/2004 as wet period; for the Puyango-Tumbes River basin (Figure 3, centre panels), the selected periods are Sep/2002-Aug/2007 as dry period, and Sep/2007-Aug/2012 as wet period; and for the Huancane River basin (Figure 3, right panels), we consider Sep/2006-Aug/2011 as dry period, and Sep/1999-Aug/2004 as wet period. The main hydroclimatic characteristics of each period are summarized in Table 2, and the contrasting hydroclimates between the selected periods are reflected in mean monthly runoff, mean monthly precipitation, average monthly temperature (Figure 3, top panels), and the flow duration curves (Figure 3, bottom panels). For example, in the Vilcanota River basin the wet (dry) period has aridity index (AI) and runoff ratio (RR) values of 1.00 (1.18) and 0.58 (0.49) respectively. Interestingly, AI values in the Puyango-Tumbes River basin are higher to those for the other catchments (Table 2), possibly because – along the Pacific coast – rainfall is higher in the north and decreases towards the south (Lavado Casimiro et al., 2012); additionally, temperature values are higher since the basin is close to the equatorial line, and hence there is high potential for evapotranspiration. Moreover, in the Puyango-Tumbes River basin, there is more runoff than precipitation between May and September (see Figure 3, upper panel), which suggests that the basin also receives groundwater contributions. Indeed, Núñez Juárez & Zegarra Loo (2006) identified aquifers located in the alluvial deposits of the Tumbes River and in the areas of the main streams, which are constantly recharged by the seasonal rains that occur in the upper part of the basin.

[Insert Figure 2]

[Insert Table 2]

[Insert Figure 3]

3.2 Hydrological modelling

In this study, we use the Framework for Understanding Structural Error (FUSE; Clark et al., 2008), which allows the implementation of an ensemble of hydrologic model structures that can be used to characterize structural uncertainty (e.g., Clark et al., 2011). Further, its modular functionality allows to diagnose inter-model differences in simulated states and fluxes through controlled experiments. FUSE discretizes the soil column along the vertical axis into an unsaturated zone (above the water table) and a saturated zone (below the water table). The major model-building decisions can be defined by the user, including the architecture of the upper and lower soil layers, and the parameterizations for simulating evaporation, surface runoff, percolation of water fluxes between soil layers, interflow, and baseflow. The multiple options available for each model building decision (Table 3) are drawn from four conceptual parent models: PRMS (Leavesley et al., 1984), Sacramento (Burnash, 1995), TOMODEL (Beven & Kirkby, 1979), and ARNO/VIC (Zhao, 1977).

It is important to note that FUSE does not perform any surface energy balance calculations, and neither does represent the canopy interception or the transpiration and evaporation from intercepted water (Clark et al., 2008). Snow is modelled using a Snow-17-based temperature index model (Anderson, 2006), which tracks snow water equivalent (SWE) based on precipitation and melt (see Henn et al., 2015 for further details). Despite all these simplifications, the models are designed to provide a robust representation of the major hydrological fluxes in the subsurface (Clark et al., 2008), and their low data requirements makes them well suited for hydrological simulations in data-scarce regions.

[Insert Table 3]

In this work, we configure 78 model structures for each basin, using 100-m elevation bands in the snow model to account for topographic effects. Model simulations are conducted at a daily time step, requiring precipitation, temperature and potential evapotranspiration (PET) – computed with the formulation proposed by Oudin et al. (2005) – as meteorological forcings. All model structures are calibrated with the Shuffled Complex Evolution (SCE-UA; Duan et al. 1992) optimization algorithm, with a maximum of 10,000 iterations, to minimize the root mean squared error (RMSE) of simulated daily streamflow:

$$\mathrm{RMSE} = \sqrt{\frac{1}{\mathrm{N}}\sum_{\mathrm{i}=1}^{\mathrm{N}}(Q_s^i - Q_o^i)^2} \qquad (1)$$

where Q_s^i and Q_o^i correspond to simulated and observed runoff, respectively, and N is the number of days in the calibration period. It should be noted that the choice of objective function is justified by the particular interest to configure models that are capable to simulate floods (i.e., peak flows) under contrasting climate scenarios. For each hydrologic model structure and each basin, two separate calibrations are conducted (i.e., one for the dry and one for the wet period). The resulting combinations of model structures and parameters – obtained for dry (MM0-dry) and wet (MM0-wet) periods – serve as the basis for the model sampling framework described below.

3.3 Selection of calibrated model structures

To sample the model space, we adopt the Pareto approach proposed by Fowler et al. (2018a) to obtain temporally stable parameter sets for a given model structure, using performance acceptance thresholds in wet and dry hydroclimatic periods. Our study differs in that we focus on the temporal transferability of model structures under a common calibration framework, rather than model parameter sets for a fixed model structure. Therefore, we apply the Pareto framework on MM0-dry and MM0-wet using five commonly used objective functions (Table 4), and considering 0.7 as the performance acceptance threshold. A small ensemble of six models is subsequently selected from the original 78-member ensembles (MM0-dry and MM0-wet) based on the following criteria:

1. Model structure and parameter set with the best performance in terms of RMSE in the calibration period (lowest RMSE Cal).
2. Model structure and parameter set with the highest NSE (or, equivalently, smallest RMSE) in both calibration and evaluation periods (highest NSE Cal-Eval).
3. Model structure and parameter set with the highest KGE in both calibration and evaluation period (highest KGE Cal-Eval).
4. Model structure and parameter set with the highest split-KGE in both calibration and evaluation periods (highest splitKGE Cal-Eval).
5. Model structure and parameter set with the highest NSE-log in both calibration and evaluation periods (highest NSElog Cal-Eval).
6. Model structure and parameter set with the highest Aggregate Objective Function (AOF; Fowler et al. 2016) in both calibration and evaluation periods (highest AOF Cal-Eval).

For criteria 2-6, we select the model structure and parameter set that meet the following requirements:

i. Efficiency indices equal or higher than 0.7 in both calibration and evaluation periods (light blue region in Figure 2).

ii. Yield the shortest Euclidean Distance (ED) with respect to the yellow point in Figure 2 (panel 3), defined as:

$$\mathrm{ED} = \sqrt{(1-\mathrm{E_{cal}})^2 + (1-\mathrm{E_{eval}})^2} \qquad (2)$$

where $\mathrm{E_{cal}}$ and $\mathrm{E_{eval}}$ are the performance metrics in calibration and evaluation periods, respectively.

In cases where condition (i) is not met by any combination of model and parameter set, we select the model structure that satisfies requirement (ii).

It should be noted that, when starting with MM0-dry (MM0-wet), the calibration period is the selected dry (wet) period. Additionally, alternative 1 does not rely on a split sample evaluation and is still common practice in hydrology. Conversely, alternatives 2-6 consider performance during both calibration and evaluation periods, and do not involve any re-calibration of model structures, but a model sampling based on evaluations with parameter values obtained from the calibration process (section 3.1.2). For example, the dry period calibration KGE (alternative 3) is the Kling-Gupta efficiency obtained with the parameters that result from model calibration conducted in the dry period, minimizing RMSE; similarly, the wet period evaluation KGE is the Kling-Gupta efficiency obtained with the same parameter set, applied in the wet period. The result of this process (i.e., application of criteria 2-6) are small (5-member) multi-model ensembles for a dry (MMP-dry) and a wet (MMP-wet) period, which provide combinations of flood-oriented models/parameters for (1) hydrologically consistent simulations, regardless of the climatic conditions, and (2) reducing the structural uncertainty in hydrologic sensitivities to climate change.

[Insert Table 4]

3.4 Hydrological consistency and inter-model agreement

We use six signature measures of hydrologic behaviour (Yilmaz et al. 2008; Pokhrel et al. 2012; Mendoza et al. 2015) to assess the model capability to reproduce the water balance, runoff seasonality and hydrological signatures from the daily flow duration curve (FDC) (see details in Appendix A). Further, we evaluate the capability of the model sampling approach to select model structures that produce hydrologically coherent simulations through the examination of monthly states and fluxes (i.e. streamflow, ET, soil moisture, snow water equivalent – SWE –, baseflow and surface runoff). Inter-model agreement in the simulation of seasonal cycles is quantified with the average standard deviation of simulated monthly variables:

$$\mathrm{Sd} = \frac{1}{12}\sum_{m=1}^{12}\sqrt{\frac{1}{n-1}\sum_{i=1}^{n}\left(x_{m,i} - \overline{x_m}\right)^2} \qquad (3)$$

Where n is the number of sample elements (i.e., number of model structures), $x_{m,i}$ is the mean monthly value of variable **x** obtained from model **i** and month **m**, and $\overline{x_m}$ is the result of averaging $x_{m,i}$ across all models.

Once the behaviour of internal model states and fluxes in MMP-dry and MMP-wet is analysed, we conduct a final screening procedure to discard potentially problematic model structures, obtaining final multi-model ensembles MMPS-dry and MMPS-wet from dry and wet calibration periods, respectively.

3.5 Hydrologic sensitivities of runoff

We compute hydrologic sensitivities for the period Sep/1986 – Aug/2016 to include both dry and wet periods. Following Vano et al. (2012), we create modified climates by using multiplicative perturbations in precipitation (70%, 80%, 90%, 100% and 110%), which are used to compute precipitation elasticities (ε), and additive perturbations in temperature (0°, 1°, 2°, 3°C) that are used for temperature sensitivities (S). Hydrologic sensitivities are estimated using 1% and 0.1°C incremental changes in precipitation and temperature, respectively, relative to each reference climate. We select these increments to be as small as possible to approximate the tangent (versus the secant), and reduce computational artifacts. We estimate ε as the fractional change in average annual runoff (Q) divided by the fractional change imposed on precipitation:

$$\varepsilon = \frac{\frac{Q_{ref+\Delta\%} - Q_{ref}}{Q_{ref}}}{\Delta\%} \qquad (4)$$

where Δ=1%. Temperature sensitivities are estimated by perturbing air temperature instead of PET, since the former variable is the most widely used in climate change impact assessments. We estimate S as the percent change in average annual runoff due to temperature changes as:

$$S = \frac{\frac{Q_{ref+\Delta} - Q_{ref}}{Q_{ref}}}{\Delta} \qquad (5)$$

where Δ=1°C.

4. RESULTS AND DISCUSSION

4.1 Choice of hydrological model structure

Figure 4 shows the coverage results from all calibrated hydrological models. In general, we observe that the Vilcanota River basin yields the best performance in contrasting climates, compared to the Huancane and Puyango-Tumbes River basins, where only a few model structures meet the acceptable performance limits. Indeed, most model structures in the Vilcanota River basin (Figure 4, left panels) are in the blue-shaded area, except when the Pareto scheme is applied with the NSE-log. For this criterion, only a few models meet the acceptance thresholds in both dry and wet periods, since the calibration objective function (RMSE) is focused on high flows; notably, all the model structures selected in the two calibration periods (coloured dots) are in the shaded area, except one structure in the NSE-log diagram for the wet calibration period (orange dot, split-KGE). In the Puyango-Tumbes River basin (Figure 4, centre panels) only a few model structures fall within the shaded area, with the lowest performances obtained when the Pareto scheme is applied with the NSE-log and AOF. However, most of the selected models (coloured dots) are in the shaded area.

Interestingly, the Huancane River basin (Figure 4, right panels) emerges as a challenging case study, since very few model structures fall within the shaded area, with the lowest performances obtained for NSE-log. In the diagrams for the dry calibration period, the model selected with NSE provides the most consistent results in contrasting climates. However, when the calibration is conducted during a wet period, few models fall within the shaded area, with evaluation metrics for which no model structures meet the acceptance thresholds (i.e., NSE, KGE). Overall, the temporal transferability of model performance in the Huancane River basin is quite poor, regardless of the evaluation criteria, since most model structures have good performance in the calibration period, but behave poorly in the evaluation period. In other words, the parameter sets found from the optimization of RMSE provide acceptable performance in terms of other evaluation metrics within the calibration period (especially if this is wet), but not necessarily for the evaluation period. This can be explained by the simplifications adopted for some process representations (e.g., canopy interception, transpiration and others), and possibly by the level of spatial disaggregation of the model structures considered here.

[Insert Figure 4]

4.2 Hydrological consistency and inter-model differences

Figures 5 and 6 illustrate the performance of all model structures, for each basin and calibration period, in terms of signature measures of hydrologic behaviour and flow duration curve, respectively. In general, the model sampling approach provides an ensemble of structures with good performance, improving inter-model agreement in comparison to the original ensemble (grey symbols in Figure 5, and gray lines in Figure 6). Further, it is observed that model performance depends considerably on the calibration period (i.e., dry or wet).

Figure 5 shows that high performance in RR, CTR y FHV signatures and low performance in FMS, FLV, and FMM signatures is obtained in all basins, suggesting that many of the structures are unable to represent low and medium flows, which is also reflected in the flow duration curve (Figure 6). It should be noted that, according to the flow duration curve graphs (Figure 6), there is an underestimation of observed low flows while, according to the hydrological signatures (Figure 5), there is an overestimation. Such discrepancy can be explained by the logarithmic transformation performed on flow values to compute some metrics (FMS, FLV and FMM; see Appendix A). Further, the poor performance of FMS, FLV and FMM signatures can be explained by the choice of RMSE as the objective calibration function and to the selected calibration period (i.e., dry or wet period). For example, in the Vilcanota River basin (Figure 5, left column) there is more bias in FLV with EVAL W-> D than in EVAL D-> W (Figure 6). This relative performance in wet and dry periods is also observed in the Huancane River basin (Figure 5, right column). However, in the Puyango-Tumbes River basin (Figure 5, center panels), there is more bias with EVAL D-> W, in comparison with EVAL W-> D (Figure 6).

[Insert Figure 5]

[Insert Figure 6]

Figure 7 displays climatological averages (Sep/1986-Aug/2016) of monthly state variables and fluxes for each basin and calibration period. One can note that the largest inter-model differences are obtained for soil moisture, SWE, baseflow and surface runoff, even if there is good agreement in streamflow seasonality among the model structures of the full ensemble (e.g., Vilcanota and Puyango-Tumbes River basins). The largest dispersion of simulated states and fluxes is obtained during spring and summer – where most precipitation events occur –, and the smallest dispersion occurs during fall and winter, except for ET in the Huancane River basin. Further, inter-model agreement improves among the selected model structures in comparison to the full 78-member ensemble provided by FUSE. For example, the multi-model selection scheme in the Vilcanota River basin for the dry period enables a considerable reduction in the standard deviation of soil moisture from 291 mm to 187 mm (35.7%) with respect to the full model ensemble (gray lines); yet, such reduction is not achieved when the model structures are calibrated in a wet period, since here the selected model structures show an increase in the standard deviation of soil moisture from 200 mm to 298 mm (49%), with respect to all model structures. In the Huancane River basin, the opposite behaviour is observed: the small multi-model reduces the original spread in simulated soil moisture when the calibration period is dry, but slightly increases the standard deviation if the models are calibrated in a wet period.

[Insert Figure 7]

The above analyses not only illustrate the potential of the model selection framework to reduce structural uncertainty in internal fluxes and states, but also highlight the need to examine simulated internal fluxes and states to discard problematic model structures. Hence, we screen the selected models based on acceptance performance thresholds, and poor or abnormal behaviour of internal states and fluxes (Figure 7). The results obtained after this model screening procedure are shown in Tables 5, 6 and 7. Note that the model structures are named differently than in other papers (i.e., FUSE 1,2,3,..,78).

In the Vilcanota River basin, we dismiss FUSE 25 and FUSE 17 from the dry calibration period (MMP-dry) and FUSE 44, FUSE 21, FUSE 01 y FUSE 61 from the wet calibration period (MMP-wet), both due to an increasing trend in the SWE daily time series (not shown here) that produced the abnormal behaviour of monthly SWE (Figure 7, left panels). In the Puyango-Tumbes River basin, we dismiss FUSE 44, FUSE 56, FUSE 59 and FUSE 62 from MMP-wet, because they do not meet the minimum performance threshold (Figure 4, center panels). For the Huancane River basin, we dismiss FUSE 03, FUSE 23, FUSE 16, FUSE 45 and FUSE 23 from MMP-dry because they do not meet the minimum performance threshold criteria. For the same basin, we discard FUSE 43, FUSE 69, FUSE 44 and FUSE 23 from MMP-wet, because they do not meet the minimum performance threshold. In the three basins, we found that none model structures from MMP-wet have been able to meet the requirements of the screening procedure to discard models (MMPS-wet). In contrast, in the model structures from MMP-dry, we found that at least one model structure has passed the screening procedure (MMPS-dry, Vilcanota and Huancane River basins). This result shows that it is possible to obtain consistent simulations of states and fluxes when the model structures are calibrated in dry periods. For example, in the Huancane River basin, only the model structure (FUSE 77) from the dry calibration period was able to pass the screening procedure. In the same basin, we discard the model structure selected with the criterion of minimizing RMSE in MMP-dry and MMP-wet, suggesting that RMSE would not be a good calibration function here.

Interestingly, we found that some model components selected before the discarding process depend on both calibration period and climatic characteristics (Tables 5, 6 and 7). For example, the Percolation (PE) component during the dry calibration period in the Vilcanota River basin is mainly represented by the PRMS equation, but when the calibration period is wet, the main modelling choice for PE (ARNO/VIC or PRMS) depends on the evaluation metric selected to apply the Pareto scheme. In the Puyango-Tumbes River basin (with high temperatures and low mean elevation, Table 1), we show that the upper layer (U) and percolation (PE) components are fully represented by ARNO/VIC - TOPMODEL and ARNO/VIC equations, respectively, regardless of the calibration period or the performance evaluation metric used when applying the Pareto scheme. For the remaining components, it is observed that these depend on the calibration period and the performance metric used when applying the Pareto scheme: for example, the TOPMODEL equation is preferred for the Surface Runoff components (SR) when the calibration period is dry, and the ARNO/VIC and TOPMODEL equation is preferred when the calibration period is wet. Finally, we do not find a preferred modelling choice for any component across MMP-dry and MMP-wet in this catchment, with the exception of the upper layer component, which is mostly represented by ARNO/VIC - TOPMODEL.

[Insert Table 5]

[Insert Table 6]

[Insert Table 7]

4.3 Hydrologic sensitivities.

Figure 8 illustrates the effects of changing precipitation and temperature on monthly streamflow values simulated by MM0-dry (gray lines) and MMP-dry (colored lines) in the Vilcanota River basin. Precipitation increments lead to increases in Q and ET under the assumption that temperature remains unchanged, and these variations are more noticeable during spring and summer, when the highest values occur. Assuming that precipitation does not change, temperature increments lead to a slight increase in ET and decrease in Q. Similar sensitivities are observed in the other basins, regardless of the calibration period (not shown).

[Insert Figure 8]

[Insert Figure 9]

We assess the spread in precipitation elasticities (ε) and temperature sensitivities (S) arising from model structure in each basin, calibration period and reference climate (Figure 9). In agreement with the results reported by Vano et al. (2012), precipitation elasticities from all model structures (Figure 9, top row) are non-linear and depend on the reference climate, with higher elasticities for drier conditions (i.e., -10%, -20% and -30%). Figure 9 (middle row) shows that lower (higher) values of mean annual streamflow are related to larger (smaller) precipitation elasticities, and higher values of annual average streamflow are related to lower values of elasticity. The simulated temperature sensitivities (S, Figure 9, bottom row) are largely negative, with the exception of some model structures in the Vilcanota River basin, since as T increases, ET increases and Q decreases in agreement with the results reported by Vano et al. (2012) for the Colorado River Basin, USA. Moreover, we observe large inter-model differences in temperature sensitivities, and no clear relationships (i.e., trends) between S values and temperature perturbations.

Figure 9 shows that, among the three catchments, the Huancane River basin provides the highest ε and S values, which in turn leads to a larger spread arising from model structures. On the other hand, the lowest ε and S values are obtained at the Puyango-Tumbes River basin, which can be explained by its geographic location (very close to the equatorial line), where higher temperatures (sea monthly averages in Figure 3) and mean annual precipitation (Table 1) are observed; in this area, small temperature (0.1°C) or precipitation (1%) perturbations do not have much impact on streamflow in comparison to the other basins.

In general, we observe that the large dispersion in precipitation elasticities and temperature sensitivities arising from the original FUSE multi-model ensemble (MM0) decrease when a Pareto scheme is applied (MMP), for all basins and calibration periods. Further, a greater reduction in structural uncertainty is obtained after conducting a model screening step based on the examination of hydrological signatures and model states and fluxes (MMPS). For example, the ensemble spread in ε and S decreases by 100% (i.e., 0.2 to 0) and 0% (i.e., 0.3 to 0.3), respectively, in the Puyango-Tumbes basin for a dry calibration period. Larger reductions are obtained in the Vilcanota and Huancane River basins in the dry calibration period (MMPS-dry), where only one model structure remains. Finally, the uncertainty in hydrologic sensitivities is greatly reduced by applying this Pareto scheme (MMP) in the three selected basins of the three hydrographic regions, highlighting the potential of the proposed approach.

5. CONCLUSIONS

In this paper, we tested the capability of a simple framework to sample hydrological model structures in order to (1) provide hydrologically consistent simulations under contrasting climatic conditions, and (2) reduce the uncertainty arising from hydrologic model choice in precipitation elasticities and temperature sensitivities. The analyses were conducted in three case study basins in Peru, representative of different hydroclimatic regimes and susceptible to flood occurrence. We configured and calibrated 78 FUSE models in dry and wet periods, obtained a sample of model structures using a Pareto scheme, refined the selection based on model diagnostics, assessed hydrological consistency and quantified hydrologic sensitivities. Further, we examined possible similarities between the selected model structures. The main conclusions are as follows:

- The proposed approach enables the identification of structures that robustly simulate catchment-scale hydrology under different climatic conditions. These models provide coherent characterizations of seasonal water balances, and perform well for various efficiency metrics and hydrological signatures that were not used in the model selection process.
- Some model components from the selected structures can be related to the climate during the calibration period and hydroclimatic characteristics of the basin. Yet,

these links are lacking for other model components, in particular in the Huancane River basin.

- The model selection procedure led to a significant reduction in the spread in precipitation elasticities and temperature sensitivities when compared to the full, 78-member model ensemble. Further, by discarding the model structures that do not meet the minimum performance thresholds and/or lead to incoherent states and fluxes, we obtained an even larger reduction in the spread of precipitation elasticities and temperature sensitivities.
- For the basins analysed here, using dry periods for model calibration and selection enhanced the robustness of simulated states and fluxes, compared to calibrations performed under wet conditions.

The results presented here reinforce the idea that inter-model agreement in climate impact metrics does not necessarily improve if traditional objective functions are used for parameter estimation and model selection (Mendoza et al., 2015). We illustrate that model evaluation under contrasting climatic conditions, together with assessments of hydrological consistency, can inform the selection of hydrological models for climate impact studies. Further, our results highlight the challenge of designing model sampling strategies that provide a coherent model ensemble in terms of process representations, especially in catchments that are 'problematic' (e.g., Huancane River basin). Future studies could address this problem by using additional sources of information (Nijzink et al., 2018; Nemri & Kinnard, 2020; Sleziak et al., 2020; Széles et al., 2020) that can be incorporated in the Pareto scheme to find behavioural combinations of model structures and parameters

ACKNOWLEDGEMENTS

Danny Saavedra acknowledges support from the National Scholarship and Educational Loan Program (PRONABEC), Peru. Pablo A. Mendoza received support from Fondecyt Project 11200142.

DATA AVAILABILITY

Daily streamflow data used in this study can be obtained from SENAMHI website https://www.senamhi.gob.pe/?&p=estaciones, and the PISCO product is freely available from

636 the IRI Data Library website

637 http://iridl.ldeo.columbia.edu/SOURCES/.SENAMHI/.HSR/.PISCO/.

638

639

640

REFERENCES

Addor, N., & Melsen, L. A. (2019). Legacy, Rather Than Adequacy, Drives the Selection of Hydrological Models. *Water Resources Research*, *55*(1), 378–390. https://doi.org/10.1029/2018WR022958

Addor, Nans, Rössler, O., Köplin, N., Huss, M., Weingartner, R., & Seibert, J. (2014). Robust changes and sources of uncertainty in the projected hydrological regimes of Swiss catchments. *Water Resources Research*, *50*(10), 7541–7562. https://doi.org/10.1002/2014WR015549

Alvarez-Garreton, C., Boisier, J. P., 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. https://doi.org/10.5194/hess-25-429-2021

ANA. (2009). *Demarcación y Delimitación de las Autoridades Administrativas del Agua*. https://sinia.minam.gob.pe/modsinia/public/docs/2826.pdf

Anderson, E. A. (2006). Snow accumulation and ablation model–SNOW-17. *US National Weather Service, Silver Spring, MD*. http://www.nws.noaa.gov/oh/hrl/nwsrfs/users_manual/part2/_pdf/22snow17.pdf.

Andres, N., Vegas Galdos, F., Casimiro, W. G. S., & Zappa M. (2014). Water resources and climate change impact modelling on a daily time scale in the Peruvian Andes. *Journal des Sciences Hydrologiques*, 59, 2043-2059. http://dx.doi.org/10.1080/02626667.2013.862336.

Aybar, C., Fernández, C., Huerta, A., Lavado, W., Vega, F., & Obando, O. F. (2019). Construction of a high-resolution gridded rainfall dataset for Peru from 1981 to the present day. *Hydrological Sciences Journal*, 1-16. https://doi.org/10.1080/02626667.2019.1649411

Bae, D.-H., Jung, I.-W., & Lettenmaier, D. P. (2011). Hydrologic uncertainties in climate change from IPCC AR4 GCM simulations of the Chungju Basin, Korea. *Journal of Hydrology*, *401*(1–2), 90–105. https://doi.org/10.1016/j.jhydrol.2011.02.012

Beck, H. E., van Dijk, A. I., De Roo, A., Miralles, D. G., McVicar, T. R., Schellekens, J., & Bruijnzeel, L. A. (2016). Global-scale regionalization of hydrologic model parameters. *Water Resources Research*, *52*(5), 3599-3622. https://doi.org/10.1002/2015WR018247

Beven, K., & Kirkby, M. (1979). A physically based, variable contributing area model of basin hydrology. *Hydrological Sciences Bulletin*., 24, 43-69. doi:10.1080/02626667909491834

Bhardwaj, K., Shah, D., Aadhar, S., & Mishra, V. (2020). Propagation of meteorological to hydrological droughts in India. *Journal of Geophysical Research: Atmospheres*, *125*(22), e2020JD033455. https://doi.org/10.1029/2020JD033455

Brigode, P., Oudin, L., & Perrin, C. (2013). Hydrological model parameter instability: A source of additional uncertainty in estimating the hydrological impacts of climate change?. *Journal of Hydrology*, *476*, 410-425. https://doi.org/10.1016/j.jhydrol.2012.11.012

Burnash, R. J. C. (1995). The NWS river forecast system-catchment modeling. *Computer models of watershed hydrology*, 311-366.

Chegwidden, O. S., Nijssen, B., Rupp, D. E., Arnold, J. R., Clark, M. P., Hamman, J. J., 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*, 623–637. https://doi.org/10.1029/2018EF001047

Chen, J., Brissette, F. P., Poulin, A., & Leconte, R. (2011). Overall uncertainty study of the hydrological impacts of climate change for a Canadian watershed. *Water Resources Research*, *47*, W12509. https://doi.org/10.1029/2011WR010602

Clark, M. P., Slater, A. G., Rupp, D. E., Woods, R. A., Vrugt, J. A., Gupta, H. V., et al. (2008). Framework for Understanding Structural Errors (FUSE): A modular framework to diagnose differences between hydrological models. *Water Resources Research*, *44*, W00B02. https://doi.org/10.1029/2007WR006735

Clark, M. P., Kavetski, D., & Fenicia, F. (2011). Pursuing the method of multiple working hypotheses for hydrological modeling. *Water Resources Research*, *47*, W09301. https://doi.org/10.1029/2010WR009827

Clark, M. P., Nijssen, B., Lundquist, J. D., Kavetski, D., Rupp, D. E., Woods, R. A., et al. (2015). A unified approach for process-based hydrologic modeling: 1. Modeling concept. *Water Resources Research*. https://doi.org/10.1002/2015WR017198

Clark, M. P., Wilby, R. L., Gutmann, E. D., Vano, J. A., Gangopadhyay, S., Wood, A. W., ... & Brekke, L. D. (2016). Characterizing uncertainty of the hydrologic impacts of climate change. *Current Climate Change Reports*, *2*(2), 55-64. DOI 10.1007/s40641-016-0034-x

Coron, L., Andréassian, V., Perrin, C., 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*, W05552. https://doi.org/10.1029/2011WR011721

Correa, S. W., de Paiva, R. C. D., Espinoza, J. C., & Collischonn, W. (2017). Multi-decadal Hydrological Retrospective: Case study of Amazon floods and droughts. *Journal of Hydrology*, *549*, 667-684. https://doi.org/10.1016/j.jhydrol.2017.04.019

Coxon, G., Freer, J., Lane, R., Dunne, T., Knoben, W. J. M., Howden, N. J. K., et al. (2019). DECIPHeR v1: Dynamic fluxEs and ConnectIvity for Predictions of HydRology. *Geoscientific Model Development*, *12*(6), 2285–2306. https://doi.org/10.5194/gmd-12-2285-2019

Craig, J. R., Brown, G., Chlumsky, R., Jenkinson, R. W., Jost, G., Lee, K., et al. (2020). Flexible watershed simulation with the Raven hydrological modelling framework. *Environmental Modelling and Software*, *129*(December 2019), 104728. https://doi.org/10.1016/j.envsoft.2020.104728

Duan, Q., Sorooshian, S., & Gupta, V. (1992). Effective and efficient global optimization for conceptual rainfall-runoff models. *Water resources research*, *28*(4), 1015-1031. https://doi.org/10.1029/91WR02985

Duethmann, D., Blöschl, G., & Parajka, J. (2020). 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. https://doi.org/10.5194/hess-24-3493-2020

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

Fowler, K., Peel, M., Western, A., & Zhang, L. (2018b). Improved rainfall-runoff calibration for drying climate: Choice of objective function. *Water Resources Research*, *54*(5), 3392-3408. https://doi.org/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. https://doi.org/10.1002/2015WR018068

Gavrilović, L., Milanović Pešić, A., & Urošev, M. (2012). A hydrological analysis of the greatest floods in Serbia in the 1960–2010 period. *Carpathian Journal of Earth and Environmental Sciences*, *7*(4), 107-116.

Grigg, A. H., & Hughes, J. D. (2018). Nonstationarity driven by multidecadal change in catchment groundwater storage: A test of modifications to a common rainfall–run-off model. *Hydrological Processes*, *32*(24), 3675–3688. https://doi.org/10.1002/hyp.13282

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

Haile, G. G., Tang, Q., Sun, S., Huang, Z., Zhang, X., & Liu, X. (2019). Droughts in East Africa: Causes, impacts and resilience. *Earth-science reviews*, *193*, 146-161. https://doi.org/10.1016/j.earscirev.2019.04.015

Heidinger, H., Carvalho, L., Jones, C., Posadas, A., & Quiroz, R. (2018). A new assessment in total and extreme rainfall trends over central and southern Peruvian Andes during 1965–2010. *International Journal of Climatology*, *38*, e998-e1015. https://doi.org/10.1002/joc.5427

Henn, B., Clark, M. P., Kavetski, D., & Lundquist, J. D. (2015). Estimating mountain basin-mean precipitation from streamflow using B ayesian inference. *Water Resources Research*, *51*(10), 8012-8033. https://doi.org/10.1002/2014WR016736

Huerta, A., Aybar, C., & Lavado-Casimiro. (2018). W. PISCO temperatura v.1.1. SENAMHI – DHI – 2018, Lima – Perú. http://iridl.ldeo.columbia.edu/documentation/.pisco/.PISCOt_report.pdf

IPCC. (2013). *Climate Change 2013: The Physical Science Basis. Contribution of Working Group I to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change*, Stocker TF et al. (eds). Cambridge University Press: Cambridge, United Kingdom and New York, NY, USA, 1535.

Jiang, T., Chen, Y. D., Xu, C., Chen, X., Chen, X., & Singh, V. P. (2007). Comparison of hydrological impacts of climate change simulated by six hydrological models in the Dongjiang Basin, South China. *Journal of Hydrology*, *336*(3–4), 316–333. https://doi.org/10.1016/j.jhydrol.2007.01.010

Kavetski, D., & Fenicia, F. (2011). Elements of a flexible approach for conceptual hydrological modeling: 2. Application and experimental insights. *Water Resources Research*, *47*(11), 1–19. https://doi.org/10.1029/2011WR010748

Klemeš, V. (1986). Operational testing of hydrological simulation models. *Hydrological Sciences Journal*, *31*(1), 13-24. https://doi.org/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) v1.0: an open-source, extendable framework providing implementations of 46 conceptual hydrologic models as continuous space-state formulations. *Geoscientific Model Development*

Discussions, 1–26. https://doi.org/10.5194/gmd-2018-332

Lavado Casimiro, W. G. S., Labat D., Guyot, J. L., & Bardin, S. A. (2011). Assessment of climate change impacts on the hydrology of the Peruvian Amazon–Andes basin. *Hydrological Processes*, 25, 3721-3734. https://doi.org/10.1002/hyp.8097

Lavado Casimiro, W. S., Ronchail, J., Labat, D., Espinoza, J. C., & Guyot, J. L. (2012). Basin-scale analysis of rainfall and runoff in Peru (1969–2004): Pacific, Titicaca and Amazonas drainages. *Hydrological Sciences Journal*, *57*(4), 625-642. https://doi.org/10.1080/02626667.2012.672985

Leavesley, G. H. (1984). *Precipitation-runoff modeling system: User's manual* (Vol. 83, No. 4238). US Department of the Interior.

Lehner, F., Wood, A. W., Vano, J. A., Lawrence, D. M., Clark, M. P., & Mankin, J. S. (2019). The potential to reduce uncertainty in regional runoff projections from climate models. *Nature Climate Change*, *9*(12), 926–933. https://doi.org/10.1038/s41558-019-0639-x

Llauca, H., Lavado-Casimiro, W., León, K., Jimenez, J., Traverso, K., & Rau, P. (2021). Assessing Near Real-Time Satellite Precipitation Products for Flood Simulations at Sub-Daily Scales in a Sparsely Gauged Watershed in Peruvian Andes. *Remote Sensing*, *13*(4), 826. https://doi.org/10.3390/rs13040826

Martínez, A., & Céspedes, L. (2017). Estudio de la vulnerabilidad presente y futura ante el cambio climático en la región Tumbes. Informe Técnico Especial.

Melsen, L. A., Addor, N., Mizukami, N., Newman, A. J., Torfs, P. J., Clark, M. P., ... & Teuling, A. J. (2018). Mapping (dis) agreement in hydrologic projections. *Hydrology and Earth System Sciences*, *22*(3), 1775-1791. https://doi.org/10.5194/hess-22-1775-2018

Mendoza, P. A., Clark, M. P., Mizukami, N., Newman, A., Barlage, M., Gutmann, E., et al. (2015). Effects of hydrologic model choice and calibration on the portrayal of climate change impacts. *Journal of Hydrometeorology*, *16*(2), 762–780. https://doi.org/10.1175/JHM-D-14-0104.1

Mendoza, P. A., Clark, M. P., Mizukami, N., Gutmann, E. D., Arnold, J. R., Brekke, L. D., & Rajagopalan, B. (2016). How do hydrologic modeling decisions affect the portrayal of climate change impacts?. *Hydrological Processes*, *30*(7), 1071-1095. https://doi.org/10.1002/hyp.10684

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

Miller, W. P., Butler, R. A., Piechota, T., Prairie, J., Grantz, K., & DeRosa, G. (2012). Water management decisions using multiple hydrologic models within the San Juan River basin under changing climate conditions. *Journal of Water Resources Planning and Management*, *138*(5), 412-420. doi:10.1061/(ASCE)WR.1943-5452.0000237

Milly, P. C., Kam, J., & Dunne, K. A. (2018). On the sensitivity of annual streamflow to air temperature. *Water Resources Research*, *54*(4), 2624-2641. https://doi.org/10.1002/2017WR021970

Mizukami, N., Clark, M. P., Gutmann, E. D., Mendoza, P. A., Newman, A. J., Nijssen, B., et al. (2016). Implications of the Methodological Choices for Hydrologic Portrayals of Climate Change over the Contiguous United States: Statistically Downscaled Forcing Data and Hydrologic Models. *Journal of Hydrometeorology*, *17*(1), 73–98. https://doi.org/10.1175/JHM-D-14-0187.1

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*, 501-515. https://doi.org/10.1016/j.jhydrol.2019.03.054

Nash, J. E., & Sutcliffe, J. V. (1970). River flow forecasting through conceptual models part I—A discussion of principles. *Journal of hydrology*, *10*(3), 282-290. https://doi.org/10.1016/0022-1694(70)90255-6

Nemri, S., & 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*, 124474. https://doi.org/10.1016/j.jhydrol.2019.124474

Nijzink, R. C., Almeida, S., Pechlivanidis, I. G., Capell, R., Gustafssons, D., Arheimer, B., ... & Hrachowitz, M. (2018). Constraining conceptual hydrological models with multiple information sources. *Water Resources Research*, *54*(10), 8332-8362. https://doi.org/10.1029/2017WR021895

Niu, G.-Y., Yang, Z.-L., Mitchell, K. E., Chen, F., Ek, M. B., Barlage, M., 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. https://doi.org/10.1029/2010JD015139

Núñez Juárez, S., & Zegarra Loo, J. (2006). Estudio geoambiental de la cuenca del río Puyango-Tumbes-[Boletín C 32]. *1560-9928*. https://hdl.handle.net/20.500.12544/264

Oudin, L., Hervieu, F., Michel, C., Perrin, C., Andréassian, V., Anctil, F., & Loumagne, C.

(2005). Which potential evapotranspiration input for a lumped rainfall–runoff model?: Part 2—Towards a simple and efficient potential evapotranspiration model for rainfall–runoff modelling. *Journal of hydrology*, *303*(1-4), 290-306. https://doi.org/10.1016/j.jhydrol.2004.08.026

Pan, Z., Liu, P., Gao, S., Xia, J., Chen, J., & Cheng, L. (2019). Improving hydrological projection performance under contrasting climatic conditions using spatial coherence through a hierarchical Bayesian regression framework. *Hydrology and Earth System Sciences*, *23*(8), 3405-3421. https://doi.org/10.5194/hess-23-3405-2019

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

Pokhrel, P., Yilmaz, K. K., & Gupta, H. V. (2012). Multiple-criteria calibration of a distributed watershed model using spatial regularization and response signatures. *Journal of Hydrology*, *418*, 49-60. https://doi.org/10.1016/j.jhydrol.2008.12.004

Pomeroy, J. W., Gray, D. M., Brown, T., Hedstrom, N. R., Quinton, W. L., Granger, R. J., & Carey, S. K. (2007). The cold regions hydrological model : a platform for basing process representation and model structure on physical evidence. *Hydrological Processes*, *21*, 2650–2667. https://doi.org/10.1002/hyp

Rau, P., Bourrel, L., Labat, D., Ruelland, D., Frappart, F., Lavado, W., ... & Felipe, O. (2018). Assessing multidecadal runoff (1970–2010) using regional hydrological modelling under data and water scarcity conditions in Peruvian Pacific catchments. *Hydrological Processes*, *33*(1), 20-35. https://doi.org/10.1002/hyp.13318

Remmers, J. O. E., Teuling, A. J., & Melsen, L. A. (2020). Can model structure families be inferred from model output? *Environmental Modelling and Software*, *133*(July), 104817. https://doi.org/10.1016/j.envsoft.2020.104817

Rivas, A. T., & Rivas, J. M. T. (2013). Estrategias de adaptación frente al cambio climático en familias rurales del altiplano puneño: estudio de caso en el centro poblado de Huancho-Huancané-Perú. *Comuni@cción: Revista de Investigación en Comunicación y Desarrollo*, *4*(1), 57-73. https://www.redalyc.org/pdf/4498/449844866006.pdf

Sanabria, J., Marengo, J., & Valverde, M. (2009). Escenarios de Cambio Climático con modelos regionales sobre el Altiplano Peruano (Departamento de Puno). Climate change scenarios using regional models for the Peruvian Altiplano (Departament of Puno). *Revista Peruana Geo-Atmosférica*, *1*, 134-149.

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

Seiller, G., Anctil, F., Perrin, C., 2012. Multimodel evaluation of twenty lumped hydrological models under contrasted climate conditions. *Hydrol. Earth Syst. Sci.* 16 (4), 1171–1189. http://dx.doi.org/10.5194/hess-16-1171-2012

Shiru, M. S., Shahid, S., Chung, E. S., & Alias, N. (2019). Changing characteristics of meteorological droughts in Nigeria during 1901–2010. *Atmospheric Research*, *223*, 60-73. https://doi.org/10.1016/j.atmosres.2019.03.010

Sleziak, P., Szolgay, J., Hlavčová, K., Danko, M., & Parajka, J. (2020). The effect of the snow weighting on the temporal stability of hydrologic model efficiency and parameters. *Journal of Hydrology*, *583*, 124639. https://doi.org/10.1016/j.jhydrol.2020.124639

Son, R., Wang, S. Y. S., Tseng, W. L., Schuler, C. W. B., Becker, E., & Yoon, J. H. (2020). Climate diagnostics of the extreme floods in Peru during early 2017. *Climate Dynamics*, *54*(1), 935-945. https://doi.org/10.1007/s00382-019-05038-y

Staudinger, M., Stahl, K., Seibert, J., Clark, M. P., & Tallaksen, L. M. (2011). Comparison of hydrological model structures based on recession and low flow simulations. *Hydrology and Earth System Sciences*, *15*(11), 3447-3459. https://doi.org/10.5194/hess-15-3447-2011

Stephens, C. M., Marshall, L. A., & Johnson, F. M. (2019). Investigating strategies to improve hydrologic model performance in a changing climate. *Journal of Hydrology*, *579*, 124219. https://doi.org/10.1016/j.jhydrol.2019.124219

Széles, B., Parajka, J., Hogan, P., Silasari, R., Pavlin, L., Strauss, P., & Blöschl, G. (2020). The added value of different data types for calibrating and testing a hydrologic model in a small catchment. *Water resources research*, *56*(10), e2019WR026153. https://doi.org/10.1029/2019WR026153

Takahashi, K., & Martínez Grimaldo, A. (2015). Impacto de la variabilidad y cambio climático en el ecosistema de Manglares de Tumbes, Perú.

Vano, J. A., Das, T., & Lettenmaier, D. P. (2012). Hydrologic sensitivities of Colorado River runoff to changes in precipitation and temperature. *Journal of Hydrometeorology*, *13*(3), 932-949. https://doi.org/10.1175/JHM-D-11-069.1

Vano, J. A., & Lettenmaier, D. P. (2014). A sensitivity-based approach to evaluating future changes in Colorado River discharge. *Climatic Change*, *122*(4), 621–634. https://doi.org/10.1007/s10584-013-1023-x

Vano, J. A., Kim, J. B., Rupp, D. E., & Mote, P. W. (2015). Selecting climate change scenarios using impact-relevant sensitivities. *Geophysical Research Letters*, *42*(13), 5516–5525. https://doi.org/10.1002/2015GL063208

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. https://doi.org/10.1016/j.jhydrol.2010.09.018

Vegas Galdos, F., Andres, N., Zappa, M., Casimiro, W. S. L., & Hilker, N. (2015), Simulation and characterization of natural regime of daily discharge in the southern Andes of Peru; Apurimac and Cusco regions. *Revista Peruana Geo-Atmosferica RPGA*, 4, 1-18. www.senamhi.gob.pe\rpga

Westra, S., Thyer, M., Leonard, M., Kavetski, D., & Lambert, M. (2014). A strategy for diagnosing and interpreting hydrological model nonstationarity. *Water Resources Research*, *50*(6), 5090–5113. https://doi.org/10.1002/2013WR014719

Wilby, R. L., & Harris, I. (2006). A framework for assessing uncertainties in climate change impacts: Low-flow scenarios for the River Thames, UK. *Water Resources Research*, *42*(2), W02419. https://doi.org/10.1029/2005WR004065

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

Zhao, R. (1977). Flood forecasting method for humid regions of China. *East China Institute of Hydraulic Engineering, Nanjing, China.*

Zulkafli, Z. D. (2014). *The hydrology of the Peruvian Amazon river and its sensitivity to climate change* (Doctoral dissertation, Imperial College London). https://ethos.bl.uk/OrderDetails.do?uin=uk.bl.ethos.650691

APPENDICES

The signature measures used here are based on formulations presented in previous climate impact studies (e.g., Mendoza et al. 2015; Yilmaz et al. 2008; Pokhrel et al. 2012). The diagnostic signature measure for water balance is the percent bias in the overall runoff ratio:

$$\%\ \text{Bias RR} = \frac{RR^s - RR^o}{RR^o} \text{x} 100 \qquad (6)$$

Where RR^o and RR^s are observed and simulated mean annual runoff ratio, respectively.

The ability to reproduce runoff seasonality is quantified by the percent bias in the centroid of the daily hydrograph for an average water year:

$$\%\ \text{Bias CTR} = \frac{\left(\sum_{j=1}^{N} t_j Q_j^s / \sum_{j=1}^{N} Q_j^s\right) - \left(\sum_{j=1}^{N} t_j Q_j^o / \sum_{j=1}^{N} Q_j^o\right)}{\left(\sum_{j=1}^{N} t_j Q_j^o / \sum_{j=1}^{N} Q_j^o\right)} \text{x} 100 \qquad (7)$$

where Q_j^o and Q_j^s are observed and simulated streamflow, respectively, at $t = t_j$, and N is the total number of days in the water year. Since the water year in Peru begins on September 1, $t_j = 1$ for that day.

The diagnostic signature measures for vertical redistribution are the percent bias in FDC mis-segment slope (% Bias FMS) and the percent bias in FDC high-segment volume (%Bias FHV). The first metric is computed as:

$$\%\ \text{Bias FMS} = \frac{\left[\log\left(Q_{m1}^s\right) - \log\left(Q_{m2}^s\right)\right] - \left[\log\left(Q_{m1}^o\right) - \log\left(Q_{m2}^o\right)\right]}{\left[\log\left(Q_{m1}^o\right) - \log\left(Q_{m2}^o\right)\right]} \text{x} 100 \qquad (8)$$

Where m1 = 0.2 and m2 = 0.7, while Q_{m1} and Q_{m2} are flows with probability of exceedance of 0.2 and 0.7, respectively. A steep slope indicates a greater flashiness in the streamflow response, while a flatter curve indicates a relatively damped response and greater storage. The percent bias in FDC high-segment volume is computed as:

$$\% \text{ Bias FHV} = \frac{\sum_{h=1}^{H}(Q_h^s - Q_h^o)}{\sum_{h=1}^{H} Q_h^o} \text{x}100 \qquad (9)$$

Where h=1,2,...,H are the flow indices in the flow matrix with probability of exceedance less than 0.02. FHV is a measure of the basin response to high precipitation and snowmelt events.

The diagnostic signature measure for long-term baseflow is the percent bias in FDC low-segment volume (%Bias FLV):

$$\% \text{ Bias FLV} = \frac{\sum_{l=1}^{L}\left[\log\left(Q_l^s\right) - \log\left(Q_L^s\right)\right] - \sum_{l=1}^{L}\left[\log\left(Q_l^o\right) - \log\left(Q_L^o\right)\right]}{\sum_{l=1}^{L}\left[\log\left(Q_l^o\right) - \log\left(Q_L^o\right)\right]} \text{x}100 \qquad (10)$$

Where l=1.2,....,L is the index within the set of values located in the FDC low flow segment (probability of exceedance between 0.7 and 1.0), and L is the index for the minimum flow.

The signature measure %Bias FMM was computed using the median value of the observed (Q_{med}^o) and simulated (Q_{med}^s) flows:

$$\% \text{ Bias FMM} = \frac{\log\left(Q_{med}^s\right) - \log\left(Q_{med}^o\right)}{\log\left(Q_{med}^o\right)} \text{x}100 \qquad (11)$$

We select the median as a measure of midrange flows, because it is less sensitive to a biased distribution than the mean of the streamflow time series.

999

TABLES

Table 1: Characteristics of the three case study watersheds. Hydroclimatic indices were computed with data from the period Sep/1986 – Aug/2016. For the Puyango-Tumbes and Huancane River basins, a shorter period was used due to the lack of streamflow records in some years.

Basin	Area [km^2]	Mean basin elevation and range [m a.s.l.]	Mean annual runoff [mm yr^{-1}]	Mean annual precipitation [mm yr^{-1}]	Mean annual PET [mm yr^{-1}]	Mean annual runoff ratio Q/P	Mean annual aridity index PET/P
Vilcanota	9586	4279 (2291-6255)	398	742	813	0.54	1.10
Puyango-Tumbes*	4694	1941 (39-3847)	718	732	1594	0.98	2.18
Huancane**	3545	4396 (3815-4976)	171	674	694	0.25	1.03

Note: PET is potential evapotranspiration; AI is the aridity index.
**Period considered: 25 years.
***Period considered: 21 years.

Table 2: Characteristics of the dry and wet periods for the three case study basins.

Basin	Period	Mean annual runoff [mm yr^{-1}]	Mean annual precipitation [mm yr^{-1}]	Mean annual PET [mm yr^{-1}]	Mean annual RR [Q/P]	Mean annual AI [PET/P]
Vilcanota	**Dry**	328	666	788	0.49	1.18
	Wet	468	806	807	0.58	1.00
Puyango-Tumbes	**Dry**	487	528	1593	0.92	3.02
	Wet	915	992	1592	0.92	1.61
Huancane	**Dry**	114	60	706	0.19	1.17
	Wet	227	750	690	0.30	0.92

1015 Table 3: FUSE model decision options (modified from Clark et al. 2008; Staudinger et al.
1016 2011).

Model structure	Model option	Existing model
Upper layer architecture "U"	Upper layer divided into tension and free storage	Sacramento
	Free storage plus tension storage sub-divided into recharge and excess	PRMS
	Upper layer defined by a single state variable	ARNO/VIC - TOPMODEL
Lower layer architecture and subsurface flow "L"	Tension storage combined with two parallel tanks	Sacramento
	Storage of unlimited size combined with linear fraction rate	PRMS
	Storage of unlimited size combined with power recession	TOPMODEL
	Storage of fixed size with non-linear storage function	ARNO/VIC
Surface runoff "SR"	ARNO/Xzang/VIC parametrization	ARNO/VIC
	PRMS variant; fraction of upper tension storage	PRMS
	TOPMODEL parametrization	TOPMODEL
Percolation "PE"	Water from field capacity to saturation available for percolation	PRMS
	Water from wilting point to saturation available for percolation	ARNO/VIC
	Percolation defined by moisture content in lower layer architecture	Sacramento

1017

1018

1019

Table 4: Performance metrics used to sample the model space produced with FUSE.

Performance metric	Equation	Range	Emphasis	References
Nash – Sutcliffe Efficiency (NSE)	$\mathrm{NSE} = 1 - \frac{\sum_{i=1}^{N}\left(Q_i^s - Q_i^o\right)^2}{\sum_{i=1}^{N}\left(Q_i^o - \overline{Q^o}\right)^2}$	-∞ – 1	High flows and dynamic discharge.	Nash & Sutcliffe (1970)
Kling – Gupta Efficiency (KGE)	$\mathrm{KGE} = 1 - \mathrm{ED};$ $\mathrm{ED} = \sqrt{(r-1)^2 + (\alpha-1)^2 + (\beta-1)^2};$ $\alpha = \frac{\sigma_S}{\sigma_O};\ \beta = \frac{\mu_S}{\mu_O};$ $r = \frac{\sum_{i=1}^{N}\left(Q_i^o - \overline{Q^o}\right)\left(Q_i^s - \overline{Q^s}\right)}{\sqrt{\sum_{i=1}^{N}\left(Q_i^o - \overline{Q^o}\right)^2}\sqrt{\sum_{i=1}^{N}\left(Q_i^s - \overline{Q^s}\right)^2}}$	-∞ – 1	Timing, streamflow variability and water balance	Gupta et al. (2009)
Split KGE	$\mathrm{Split\ KGE} = \frac{1}{T}\sum_{t=1}^{T}\mathrm{KGE}_t$	-∞ – 1	Same as KGE, but no year can have more influence than any other year	Fowler et al. (2018b)
Nash – Sutcliffe Efficiency with logarithmic transf. (NSE-log)	$\mathrm{NSElog} = 1 - \frac{\sum_{i=1}^{N}\left(\log\left(Q_i^s\right) - \log\left(Q_i^o\right)\right)^2}{\sum_{i=1}^{N}\left(\log\left(Q_i^o\right) - \log\left(\overline{Q^o}\right)\right)^2}$	-∞ – 1	Low flows.	Nash & Sutcliffe (1970) Santos et al. (2018)
Aggregate objective function (AOF)	$\mathrm{AOF} = \frac{\mathrm{AOF_{sig}} + \mathrm{AOF_{gof}}}{2}$ $\mathrm{AOF_{sig}} = 1 - \sum_{q=1}^{8}\frac{Y_{q,o} - Y_{q,s}}{8\ \sigma_q}$ $\mathrm{AOF_{gof}} = \frac{2B + r + r_{\log}}{4}$	-∞ – 1	-	Beck et al. (2016)

Note: Q_i^o is the observed daily runoff; Q_i^s is the simulated daily runoff; $\overline{Q^o}$ is the mean of the observed daily runoff values; $\overline{Q^s}$ is the mean of the simulated daily runoff values. N and T represent the total number of days and water years, respectively, used to compute efficiency metrics. B and r represent the bias and Pearson correlation coefficient, respectively, computed between simulated and observed daily runoff; and r_{log} is the Pearson correlation coefficient computed between natural-log transformed simulated and observed runoff.

Table 5: Components of the hydrological model structures obtained from the application of the Pareto scheme in the Vilcanota River basin, for both dry (MMP-dry), and wet (MMP-wet) calibration periods. The reference model structure that provides the lowest RMSE during the calibration period is included for comparison purposes, and the model structures discarded due to abnormal behavior of states and/or fluxes are in italics and bold.

Calibration Period	**Selection Criteria**	**Model structure name**	**Model structure components**			
			U	**L**	**SR**	**PE**
Dry period	smallest RMSE Cal	FUSE 01	Sacramento	Sacramento	ARNO/VIC	PRMS
	higher NSE Cal-Eval	***FUSE 25***	***PRMS***	***PRMS***	***ARNO/VIC***	***PRMS***
	higher KGE Cal-Eval	***FUSE 25***	***PRMS***	***PRMS***	***ARNO/VIC***	***PRMS***
	higher splitKGE Cal-Eval	***FUSE 25***	***PRMS***	***PRMS***	***ARNO/VIC***	***PRMS***
	higher NSElog Cal-Eval	FUSE 43	ARNO/VIC - TOPMODEL	Sacramento	ARNO/VIC	PRMS
	higher AOF Cal-Eval	***FUSE 17***	***Sacramento***	***TOPMODEL***	***TOPMODEL***	***PRMS***
Wet period	***smallest RMSE Cal***	***FUSE 44***	***ARNO/VIC - TOPMODEL***	***Sacramento***	***ARNO/VIC***	***ARNO/VIC***
	higher NSE Cal-Eval	***FUSE 44***	***ARNO/VIC - TOPMODEL***	***Sacramento***	***ARNO/VIC***	***ARNO/VIC***
	higher KGE Cal-Eval	***FUSE 44***	***ARNO/VIC - TOPMODEL***	***Sacramento***	***ARNO/VIC***	***ARNO/VIC***
	higher splitKGE Cal-Eval	***FUSE 21***	***Sacramento***	***ARNO/VIC***	***PRMS***	***PRMS***
	higher NSElog Cal-Eval	***FUSE 01***	***Sacramento***	***Sacramento***	***ARNO/VIC***	***PRMS***
	higher AOF Cal-Eval	***FUSE 61***	***ARNO/VIC - TOPMODEL***	***TOPMODEL***	***ARNO/VIC***	***PRMS***

1039 Table 6: Same as Table 5, but for the Puyango-Tumbes River basin.

Calibration Period	Selection Criteria	Model structure name	Model structure components			
			U	L	SR	PE
Dry period	smallest RMSE Cal	FUSE 50	ARNO/VIC - TOPMODEL	Sacramento	TOPMODEL	ARNO/VIC
	higher NSE Cal-Eval	FUSE 59	ARNO/VIC - TOPMODEL	PRMS	TOPMODEL	ARNO/VIC
	higher KGE Cal-Eval	FUSE 59	ARNO/VIC - TOPMODEL	PRMS	TOPMODEL	ARNO/VIC
	higher splitKGE Cal-Eval	FUSE 59	ARNO/VIC - TOPMODEL	PRMS	TOPMODEL	ARNO/VIC
	higher NSElog Cal-Eval	FUSE 62	ARNO/VIC - TOPMODEL	TOPMODEL	ARNO/VIC	ARNO/VIC
	higher AOF Cal-Eval	FUSE 65	ARNO/VIC - TOPMODEL	TOPMODEL	PRMS	ARNO/VIC
Wet period	***smallest RMSE Cal***	***FUSE 44***	***ARNO/VIC – TOPMODEL***	***Sacramento***	***ARNO/VIC***	***ARNO/VIC***
	higher NSE Cal-Eval	***FUSE 56***	***ARNO/VIC – TOPMODEL***	***PRMS***	***PRMS***	***ARNO/VIC***
	higher KGE Cal-Eval	***FUSE 59***	***ARNO/VIC – TOPMODEL***	***PRMS***	***TOPMODEL***	***ARNO/VIC***
	higher splitKGE Cal-Eval	***FUSE 59***	***ARNO/VIC – TOPMODEL***	***PRMS***	***TOPMODEL***	***ARNO/VIC***
	higher NSElog Cal-Eval	***FUSE 62***	***ARNO/VIC – TOPMODEL***	***TOPMODEL***	***ARNO/VIC***	***ARNO/VIC***
	higher AOF Cal-Eval	***FUSE 62***	***ARNO/VIC - TOPMODEL***	***TOPMODEL***	***ARNO/VIC***	***ARNO/VIC***

1040

1041

Table 7: Same as Table 5, but for the Huancane River basin.

Calibration Period	Selection Criteria	Model structure name	Model structure components			
			U	L	SR	PE
Dry period	***smallest RMSE Cal***	***FUSE 03***	***Sacramento***	***Sacramento***	***PRMS***	***PRMS***
	higher NSE Cal-Eval	FUSE 77	ARNO/VIC - TOPMODEL	ARNO/VIC	TOPMODEL	ARNO/VIC
	higher KGE Cal-Eval	***FUSE 23***	***Sacramento***	***ARNO/VIC***	***TOPMODEL***	***PRMS***
	higher splitKGE Cal-Eval	***FUSE 16***	***Sacramento***	***TOPMODEL***	***PRMS***	***Sacramento***
	higher NSElog Cal-Eval	***FUSE 45***	***ARNO/VIC – TOPMODEL***	***Sacramento***	***ARNO/VIC***	***Sacramento***
	higher AOF Cal-Eval	***FUSE 23***	***Sacramento***	***ARNO/VIC***	***TOPMODEL***	***PRMS***
Wet period	***smallest RMSE Cal***	***FUSE 43***	***ARNO/VIC – TOPMODEL***	***Sacramento***	***ARNO/VIC***	***PRMS***
	higher NSE Cal-Eval	***FUSE 69***	***ARNO/VIC – TOPMODEL***	***TOPMODEL***	***TOPMODEL***	***Sacramento***
	higher KGE Cal-Eval	***FUSE 69***	***ARNO/VIC – TOPMODEL***	***TOPMODEL***	***TOPMODEL***	***Sacramento***
	higher splitKGE Cal-Eval	***FUSE 69***	***ARNO/VIC – TOPMODEL***	***TOPMODEL***	***TOPMODEL***	***Sacramento***
	higher NSElog Cal-Eval	***FUSE 44***	***ARNO/VIC – TOPMODEL***	***Sacramento***	***ARNO/VIC***	***ARNO/VIC***
	higher AOF Cal-Eval	***FUSE 23***	***Sacramento***	***ARNO/VIC***	***TOPMODEL***	***PRMS***

FIGURE LEGENDS

Figure 1: Location and elevation of the three case study basins

Figure 2: Flowchart illustrating the methodology.

Figure 3: Catchment-averaged mean monthly values of runoff, precipitation and air temperature (top panels), and flow duration curves (bottom panels) for the selected dry (red lines) and wet (blue lines) periods. The results for each basin are displayed in different columns.

Figure 4: Coverage results from all calibrated model structures, for each basin and each calibration period (displayed along different columns). The horizontal and vertical dashed lines indicate performance acceptance thresholds, and the light blue region represents the region where temporally consistent performance is obtained. The red triangle represents the combination of model structure and parameter set that minimizes RMSE during the calibration period (i.e., the common practice); the colored dots represent the models that were selected using the criteria defined in section 3.1.3, and the remaining models are displayed as gray dots.

Figure 5: Percent biases in signature measures of hydrologic behavior (rows) for each basin and each calibration period (columns), where EVAL W->D (D->W) indicates model performance in a dry (wet) period with parameters calibrated in a wet (dry) period.

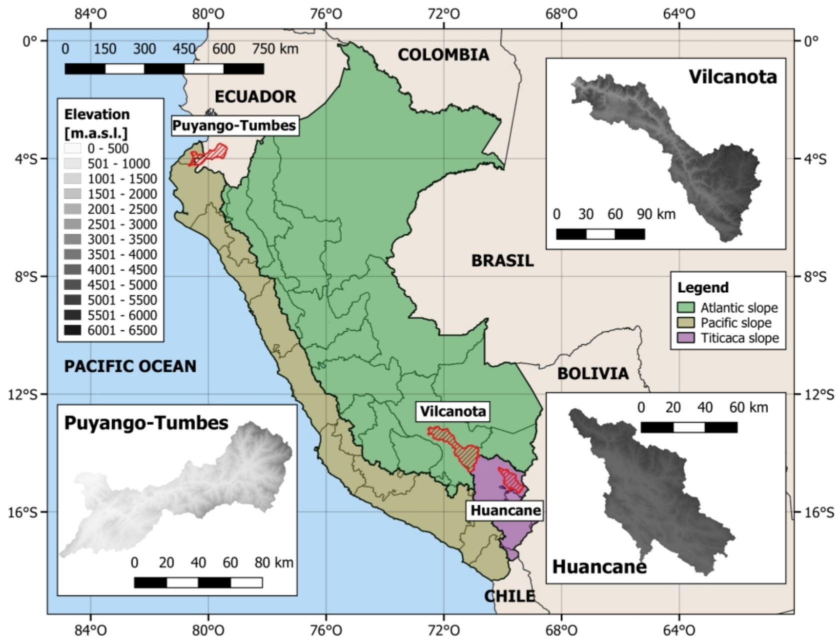
Figure 6: Flow duration curves for each basin and each calibration period. The black line represents observations, gray lines represent the full multi-model ensemble, and the model structures selected with different performance evaluation criteria are displayed in colored lines.

Figure 7: Monthly average fluxes and states for each basin and each calibration period, considering a 30-year period (September/1986 – August/2016). Inter-model agreement is quantified with an ensemble spread metric (equation 2), displayed at the top of each panel for the full ensemble (left) and the five model structures – represented by colored lines – selected with the Pareto scheme (right). The reference model structure that provides the lowest RMSE during the calibration period is displayed in dashed red, for comparison purposes. The observed average monthly streamflow values (black line, upper panel) are only shown as a reference and not for evaluation purposes, since there is not enough information available for 30 years in the Puyango-Tumbes and Huancané river basins.

Figure 8: Climatological monthly averages (September/1986 – August/2016) of runoff and ET obtained with model parameters calibrated in a dry period. Results are displayed for (top) precipitation perturbations of 70%, 80%, 90% and 110%, and (bottom) temperature increases of 1°, 2° and 3°C. The gray lines show the results with the full ensemble (MM0-dry), and the colored lines show the results obtained with the multi-model ensemble obtained from the application of the Pareto scheme (MMP). The model structures discarded during the screening procedure are plotted with x, and the accepted model structures (MMS) are plotted with circles.

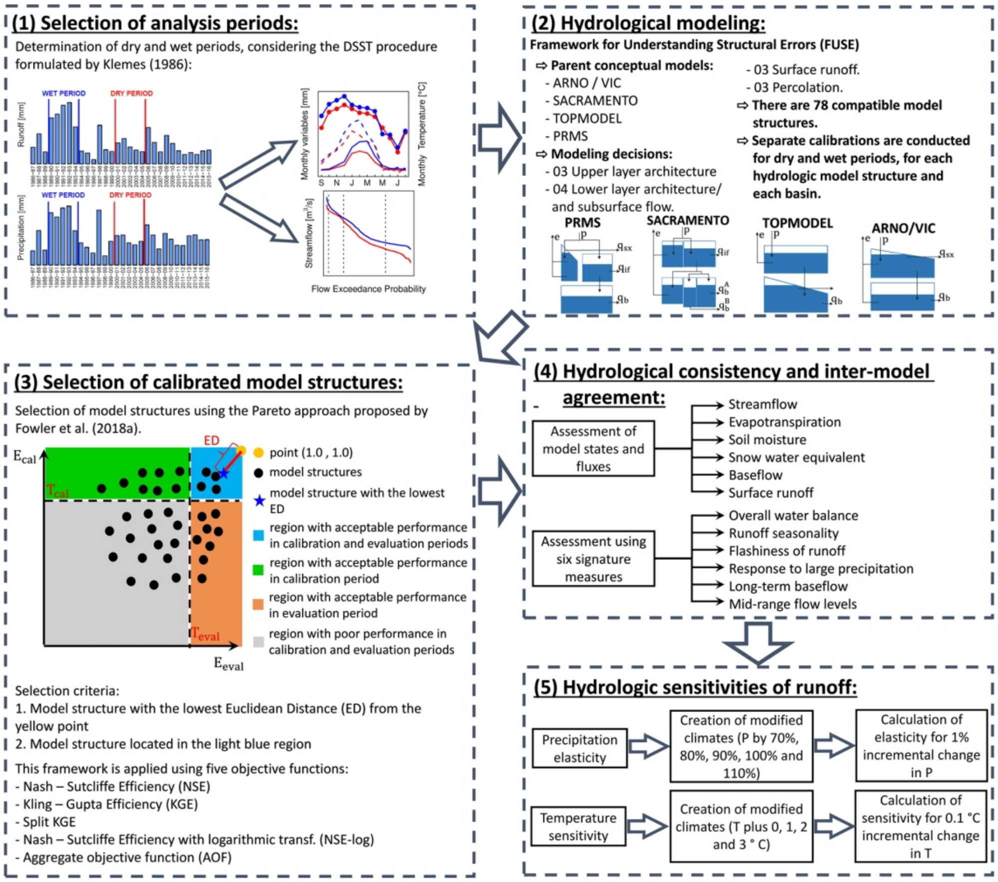
Figure 9: Precipitation elasticity (ε) and temperature sensitivities (S) for each basin and calibration period, computed for the 30-year period (September/1986 – August/2016). In the top panels, the x-axis represents the percent changes in precipitation from the reference climates; in the middle panels, the x-axis represents the changes in mean annual runoff from

the reference climates, and in the bottom panels the x-axis represents the additive temperature changes from the reference climates. The average monthly standard deviations obtained from the full ensemble (MM0) and the final multi-model ensemble (MMPS) are displayed at the top of each panel. Discarded model structures are represented with x, and accepted model structures are plotted with circles. The vertical dashed line is the observed mean annual streamflow, which is only shown as a reference and not for evaluation purposes, since there is not enough information available for 30 years in the Puyango-Tumbes and Huancané river basins.



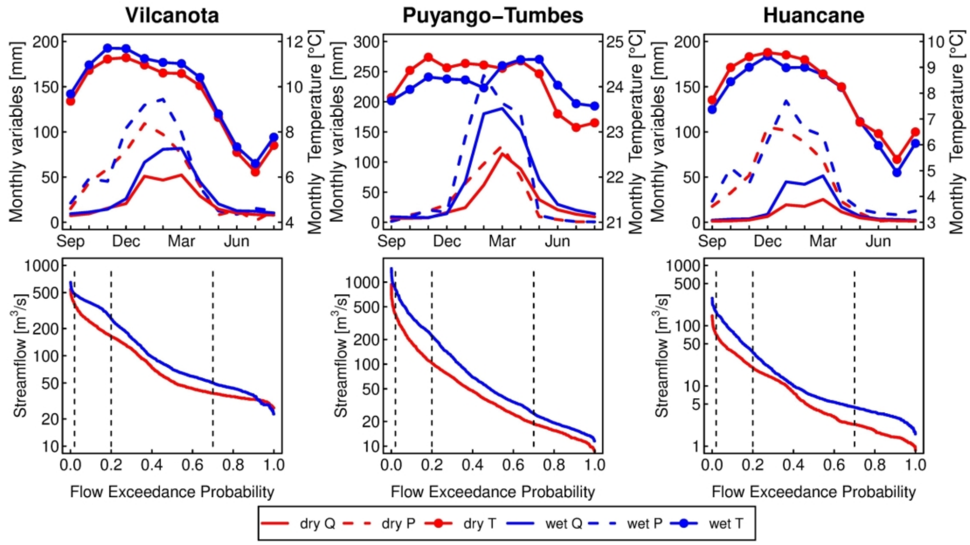
Location and elevation of the three case study basins.

825x583mm (72 x 72 DPI)



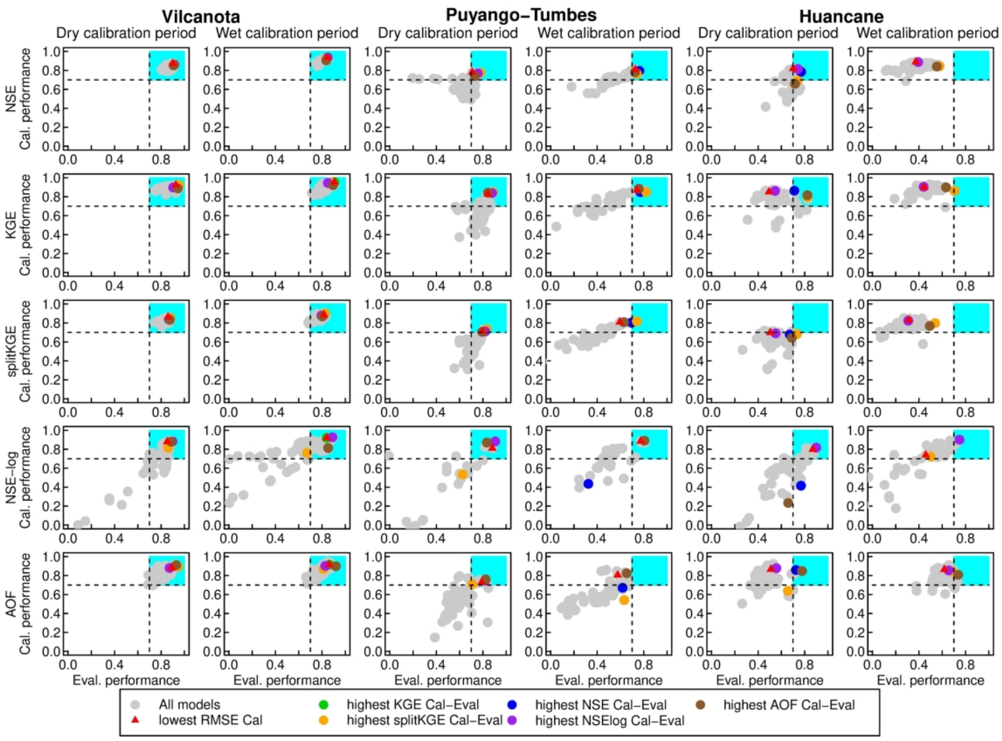
Flowchart illustrating the methodology.

1311x1155mm (72 x 72 DPI)



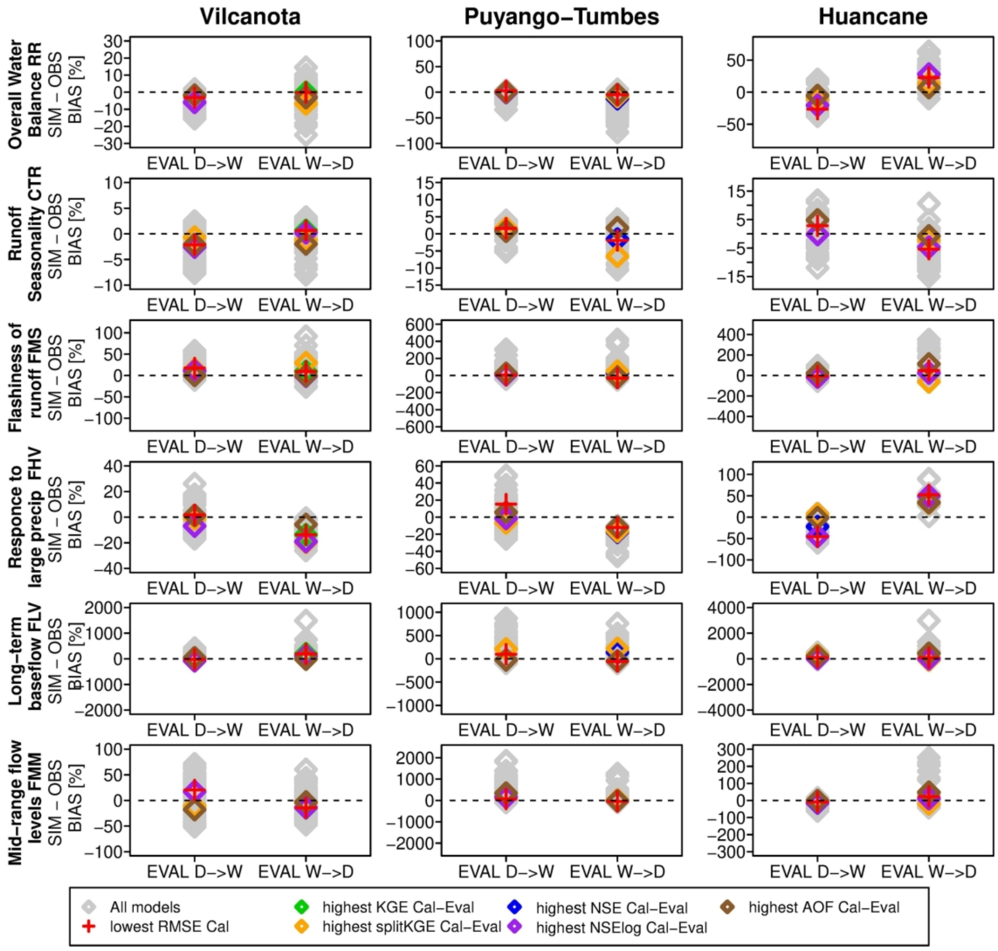
Catchment-averaged mean monthly values of runoff, precipitation and air temperature (top panels), and flow duration curves (bottom panels) for the selected dry (red lines) and wet (blue lines) periods. The results for each basin are displayed in different columns.

198x108mm (300 x 300 DPI)



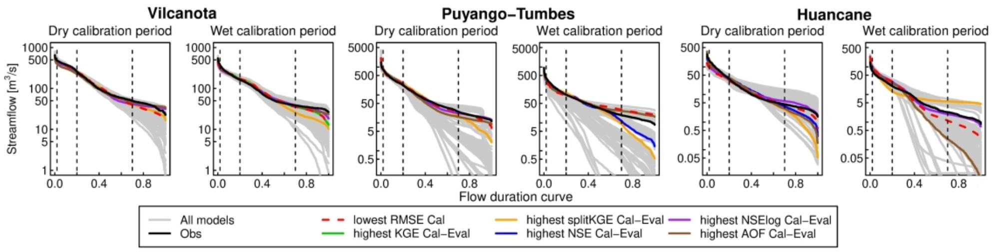
Coverage results from all calibrated model structures, for each basin and each calibration period (displayed along different columns). The horizontal and vertical dashed lines indicate performance acceptance thresholds, and the light blue region represents the region where temporally consistent performance is obtained. The red triangle represents the combination of model structure and parameter set that minimizes RMSE during the calibration period (i.e., the common practice); the colored dots represent the models that were selected using the criteria defined in section 3.1.3, and the remaining models are displayed as gray dots.

238x174mm (300 x 300 DPI)



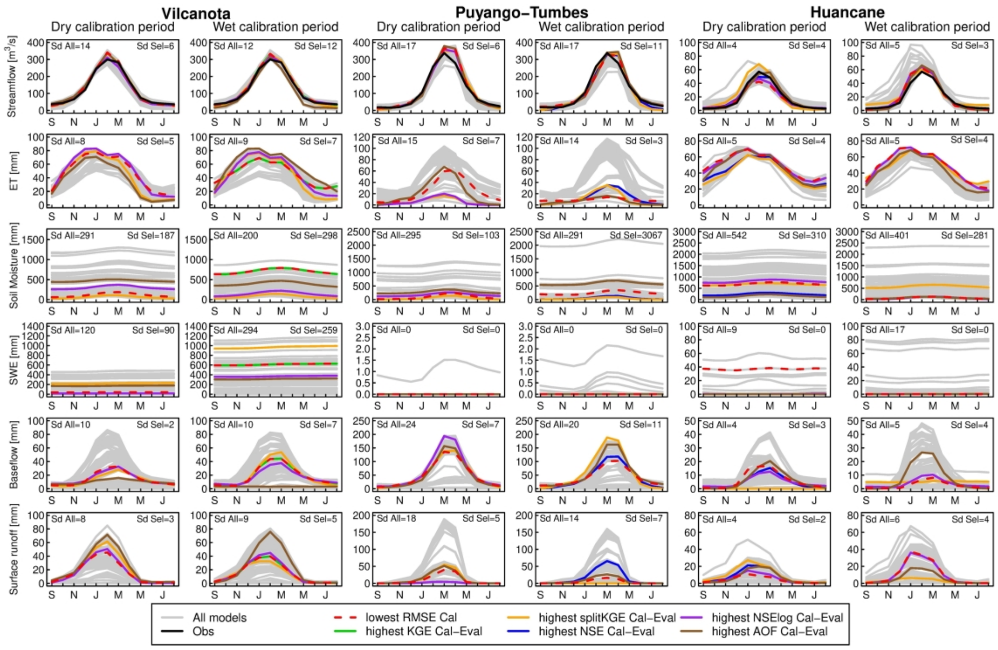
Percent biases in signature measures of hydrologic behavior (rows) for each basin and each calibration period (columns), where EVAL W->D (D->W) indicates model performance in a dry (wet) period with parameters calibrated in a wet (dry) period.

184x174mm (300 x 300 DPI)



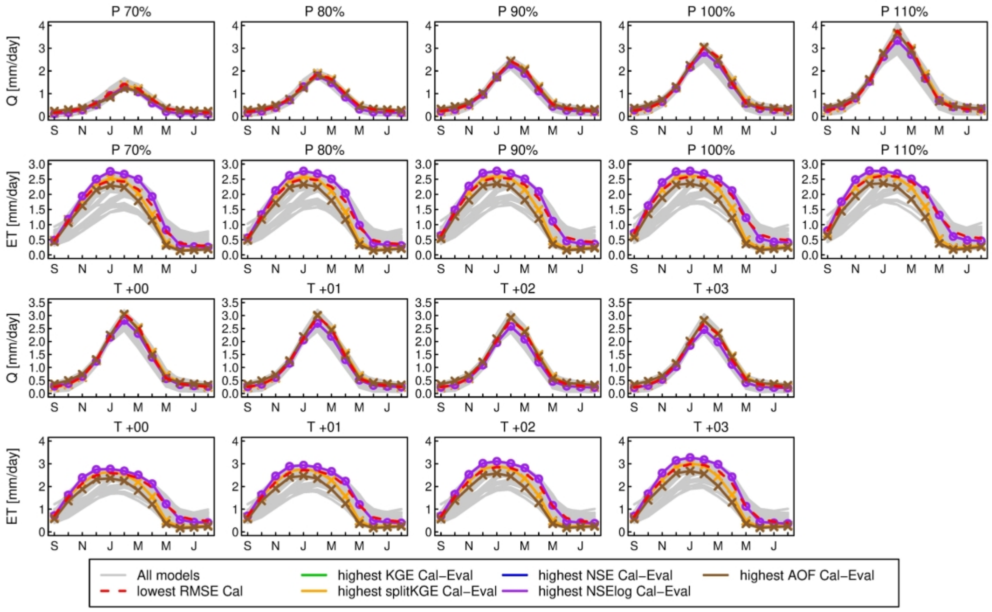
Flow duration curves for each basin and each calibration period. The black line represents observations, gray lines represent the full multi-model ensemble, and the model structures selected with different performance evaluation criteria are displayed in colored lines.

269x66mm (300 x 300 DPI)



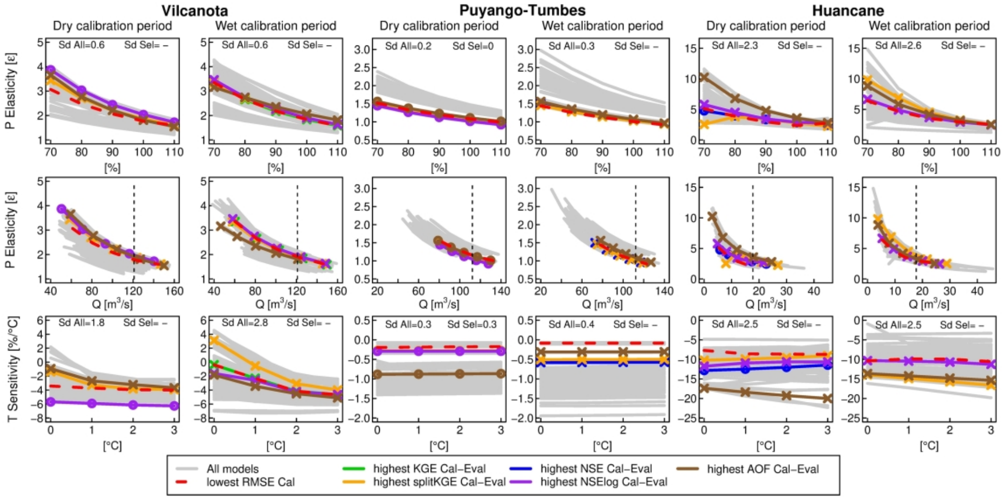
Monthly average fluxes and states for each basin and each calibration period, considering a 30-year period (September/1986 – August/2016). Inter-model agreement is quantified with an ensemble spread metric (equation 2), displayed at the top of each panel for the full ensemble (left) and the five model structures – represented by colored lines – selected with the Pareto scheme (right). The reference model structure that provides the lowest RMSE during the calibration period is displayed in dashed red, for comparison purposes. The observed average monthly streamflow values (black line, upper panel) are only shown as a reference and not for evaluation purposes, since there is not enough information available for 30 years in the Puyango-Tumbes and Huancané river basins.

269x174mm (300 x 300 DPI)



Climatological monthly averages (September/1986 – August/2016) of runoff and ET obtained with model parameters calibrated in a dry period. Results are displayed for (top) precipitation perturbations of 70%, 80%, 90% and 110%, and (bottom) temperature increases of 1°, 2° and 3°C. The gray lines show the results with the full ensemble (MM0-dry), and the colored lines show the results obtained with the multi-model ensemble obtained from the application of the Pareto scheme (MMP). The model structures discarded during the screening procedure are plotted with x, and the accepted model structures (MMS) are plotted with circles.

232x145mm (300 x 300 DPI)



Precipitation elasticity (ε) and temperature sensitivities (S) for each basin and calibration period, computed for the 30-year period (September/1986 – August/2016). In the top panels, the x-axis represents the percent changes in precipitation from the reference climates; in the middle panels, the x-axis represents the changes in mean annual runoff from the reference climates, and in the bottom panels the x-axis represents the additive temperature changes from the reference climates. The average monthly standard deviations obtained from the full ensemble (MM0) and the final multi-model ensemble (MMPS) are displayed at the top of each panel. Discarded model structures are represented with x, and accepted model structures are plotted with circles. The vertical dashed line is the observed mean annual streamflow, which is only shown as a reference and not for evaluation purposes, since there is not enough information available for 30 years in the Puyango-Tumbes and Huancané river basins.

280x138mm (300 x 300 DPI)