Reducing structural uncertainty in conceptual hydrological modeling in the semi-arid Andes

P. HUBLART[1,4], D. RUELLAND[2], A. DEZETTER[3] & H. JOURDE[1]
[1] *UM2,* [2] *CNRS,* [3] *IRD – UMR HydroSciences Montpellier, Place E. Bataillon, 34395 Montpellier Cedex 5, France*
[4] *Centro de Estudios Avanzados en Zonas Áridas (CEAZA), Raúl Bitrán s/n, La Serena, Chile*
paul.hublart@um2.fr / denis.ruelland@um2.fr

Abstract The use of lumped, conceptual models in hydrological impact studies requires placing more emphasis on the uncertainty arising from deficiencies and/or ambiguities in the model structure. This study provides an opportunity to combine a multiple-hypothesis framework with a multi-criteria assessment scheme to reduce structural uncertainty in the conceptual modeling of a meso-scale Andean catchment (1515 km²) over a 30-year period (1982–2011). The modeling process was decomposed into six model-building decisions related to the following aspects of the system behavior: snow accumulation and melt, runoff generation, redistribution and delay of water fluxes, and natural storage effects. Each of these decisions was provided with a set of alternative modeling options, resulting in a total of 72 competing model structures. These structures were calibrated using the concept of Pareto optimality with three criteria pertaining to streamflow simulations and one to the seasonal dynamics of snow processes. The results were analyzed in the four-dimensional space of performance measures using a fuzzy *c*-means clustering technique and a differential split sample test, leading to identify 14 equally acceptable model hypotheses. A filtering approach was then applied to these best-performing structures in order to minimize the overall uncertainty envelope while maximizing the number of enclosed observations. This led to retain 8 model hypotheses as a representation of the minimum structural uncertainty that could be obtained with this modeling framework. Future work to better consider model predictive uncertainty should include a proper assessment of parameter equifinality and data errors, as well as the testing of new or refined hypotheses to allow for the use of additional auxiliary observations.

1. INTRODUCTION

Conceptual catchment models based on the combination of several schematic stores are popular tools in flood forecasting and water resources management (e.g. Jakeman and Letcher, 2003; Xu and Singh, 2004). The main rationale behind this success lies in the fact that relatively simple structures with low data and computer requirements generally outweigh the performance of far more complex physically-based models (e.g. Michaud and Sorooshian, 1994; Refsgaard and Knudsen, 1996; Kokkonen and Jakeman, 2001). Also, most water management decisions are made at operational scales having much more to do with catchment-scale administrative considerations than with our understanding of fine-scale processes. As a result, conceptual models are being increasingly used to evaluate the potential impacts of climate change on hydrological systems (e.g. Minville et al., 2008; Ruelland et al., 2012) and freshwater availability (e.g. Milano et al., 2013; Collet et al., 2013).

This modeling strategy, however, is regularly criticized for oversimplifying the physics of catchments and leading to unreliable simulations when conditions shift beyond the range of prior experience. Part of the problem comes from the fact that model structures are usually specified *a priori*, based on preconceived opinions about how systems work, which in general leads to an excessive dependence on the calibration process. More than a lack of physical background, this practice reveals a misunderstanding about *how* such models should be based on physics (Kirchner, 2006; Blöschl and Montanari, 2010). Hydrological systems are not structureless things composed of randomly distributed elements, but rather self-organizing systems characterized by the emergence of macroscale patterns and structures (Dooge, 1986; Sivapalan, 2006; Ehret et al., 2014). As such, the reductionist idea that catchments can be understood by merely aggregating (upscaling) fine-scale mechanistic laws is generally misleading (Dooge, 1997; McDonnell et al., 2007). Self-organization at the catchment scale means that new hydrologic relationships with fewer degrees of freedom have to be envisioned (e.g. McMillan, 2012a). Yet, finding simplicity in complexity does not imply that simple models available in the literature can be used as ready-made engineering tools with little or no consideration for the specific features of each catchment (Wainwright and Mulligan, 2004; Savenije, 2009). As underlined by Kirchner (2006), it is important to ensure that the "right answers" are obtained for the "right reasons". In the case of poorly-defined systems where physically-oriented interpretations can only be sought *a posteriori* to check for the model realism, this requires placing more emphasis on the uncertainty arising from deficiencies and/or ambiguities in the model structure than is currently done in most hydrological impact studies.

Structural uncertainty can be described in terms of *inadequacy* and *non-uniqueness*. Model inadequacy arises from the many simplifying assumptions and epistemic errors made in the selection of which processes to represent and how to represent them. It reflects the extent to which a given model differs from the real system it is intended to represent. In practice, this results in the failure to capture all relevant aspects of the system behavior within a single model structure or parameter set. A common way of addressing this source of uncertainty is to adopt a top-down approach to model-building (Jothityangkoon et al., 2001; Sivapalan et al., 2003), in which different models of increasing complexity are tested to determine the adequate level of process representation. Where fluxes and state variables are made explicit, alternative data sources (other than streamflow) such as groundwater levels (Seibert, 2000; Seibert and McDonnell, 2002), tracer samples (Son and Sivapalan, 2007; Birkel et al., 2010; Capell et al., 2012) or snow measurements (Clark et al., 2006; Parajka and Blöschl, 2008), can also be used to improve the internal consistency of model structures. Additional criteria can then be introduced in relation to these auxiliary data or to specific aspects of the hydrograph (driven vs. nondriven components, rising limb, recession limbs...). In this perspective, multi-criteria evaluation techniques based on the concept of Pareto-optimality provide an interesting way to both reduce and quantify structural inadequacy (Gupta et al., 1998; Boyle et al., 2000; Efstratiadis and Koutsoyiannis, 2010). A parameter set is said to be Pareto-optimal if it cannot be improved upon without degrading at least one of the objective criteria. In general, meaningful information on the origin of model deficiencies can be derived from the mapping of Pareto-optimal solutions in the space of performance measures (often called the Pareto front) and used to discriminate between several rival structures (Lee et al., 2011). Further, the Pareto set of solutions obtained with a given model is commonly used to generate simulation envelopes (hereafter called 'Pareto-envelopes' for brevity's sake) representing the uncertainty associated with structural errors (i.e. model inadequacy).

Non-uniqueness refers to the existence of many different model structures (and parameter sets) giving equally acceptable fits to the observed data. Structural inadequacy and the limited (and often uncertain) information of the available data make it highly unlikely to identify a single, unambiguous representation of how a system works. There may be, for instance, many different possible representations of flow pathways yielding the same integral signal (e.g. streamflow) at the catchment outlet (Schaefliet al., 2011). Non-uniqueness in model identification has also been widely described in terms of equifinality (Beven, 1993 and 2006) and may be viewed as a special case of a more general epistemological issue known as the “underdetermination” problem. Over the past decade, these considerations have encouraged a shift in focus toward more flexible modeling tools based on the concept of multiple working hypotheses (Buytaert and Beven, 2011; Clark et al., 2011). A number of modular frameworks have been proposed, in which model components (i.e. individual hypotheses) can be assembled and connected in many ways to build a variety of alternative model structures (i.e. overall hypotheses). Recent examples of such modular modeling frameworks (MMF) include the Imperial College Rainfall-Runoff Modeling Toolbox (RRMT) (Wagener et al., 2002), the Framework for Understanding Structural Errors (FUSE) (Clark et al., 2008) and the SUPERFLEX modeling environment (Fenicia et al., 2011). Clark et al. (2011) suggested that this approach to model identification represents a valuable alternative to “most practical applications of the top-down approach”, which “seldom consider competing process representations of equivalent complexity”. Compared to current multimodel strategies, MMF also provide the possibility to better scrutinize the effect of each individual hypothesis (i.e. model component), provided that the model decomposition is sufficiently fine-grained. Finally, Clark et al. (2011) argued that ensembles of competing model structures obtained from MMF (both of equal and varying complexity) can also be used to quantify the structural uncertainty arising because of system non-identifiability (i.e. model non-uniqueness). So far, however, this method has mostly been applied to relatively small (<500 km^2) and humid catchments of the Northern Hemisphere (Krueger et al., 2010; Smith and Marshall, 2010; Staudinger et al., 2011; Kavetski and Fenicia, 2011; McMillan et al., 2012b; Coxon et al., 2013), with less attention being given to larger scales of interest (>1000 km^2) and semi-arid regions (e.g. Clark et al., 2008). Moreover, several of these studies have insisted on the need for multiple criteria related to different aspects of the system’s behavior in order to improve the usefulness of MMF. Yet, most of the time these additional criteria or signatures were not used to guide model development or constrain calibration but rather as posterior diagnostics in validation (see Kavetski and Fenicia, 2011). Thus, the potential benefits of using the concept of Pareto-efficiency to constrain model development and help differentiate between numerous competing hypotheses remain largely unexplored in the current literature devoted to MMF. Also, very few studies have included alternative conceptual representations of snow processes in their modular frameworks (e.g. Smith and Marshall, 2010), even though snowmelt may have played a significant role in several cases (Clark et al., 2008; Staudinger et al., 2011).

Addressing these issues is of particular importance in the case of arid to semi-arid Andean catchments such as those found around 30°S. The Norte Chico region of Chile, in particular, has been identified as being highly vulnerable to climate change impacts in a number of recent reports (IPCC, 2013) and studies (e.g. Souvignet et al., 2010; Young et al., 2010). Yet, very few catchments in this region have been studied intensively enough to provide reliable model simulations, often with no estimation of the surrounding uncertainty (Souvignet, 2007; Ruelland et al., 2011; Vicuña et al., 2011; Hublart et al., 2013). This study is the first step of a larger research project, whose final aim is to assess the capacity to meet current and future irrigation water requirements in a mesoscale catchment of the Norte Chico region. The objective here is to provide a set of reasonable model structures that can be used for the hydrological modeling of the catchment. To achieve this goal, a MMF was developed and combined with a multi-criteria optimization framework using streamflow and satellite-based snow cover data.

2. STUDY AREA

2.1. General site description

The Claro River Catchment is a semi-arid, mountainous catchment located in the northeastern part of the Coquimbo region, in north-central Chile (Fig. 1). It drains an area of approximately 1515 km², characterized by high elevations ranging from 820 m a.s.l. at the basin outlet (Rivadavia) to over 5500 m a.s.l. in the Andes Cordillera. The topography is dominated by a series of generally north-trending, fault-bounded mountain blocks interspersed with a few steep-sided valleys.

The underlying bedrock consists almost entirely of granitic rocks ranging in age from Pennsylvanian to Oligocene and locally weathered to saprolite. Above 3000 m a.m.s.l., repeated glaciations and the continuous action of frost and thaw throughout the year have caused an intense shattering of the exposed rocks (Caviedes and Paskoff, 1975), leaving a landscape of bare rock and screes almost devoid of soil.

The valley-fill material consists of mostly unconsolidated Quaternary alluvial sediments mantled by generally thin soils (< 1 m) of sandy to sandy-loam texture. Vineyards and orchards cover most of the valley floors and lower hill slopes but account for less than 1% of the total catchment area. Most of the annual precipitation, however, occurs as snow during the winter months, leading to an entire dependence on surface-water resources to satisfy crop water needs during the summer. Irrigation water abstractions occur at multiple locations along the river's course depending on both historical water rights and water availability. By contrast, natural vegetation outside the valleys is extremely sparse and composed mainly of subshrubs (e.g. *Adesmia echinus*) and cushion plants (e.g. *Laretia acaulis*, *Azorella compacta*) with very low transpiration rates (Squeo et al., 1993). The Claro River originates from a number of small tributaries flowing either permanently or seasonally in the mountains.

2.2. Hydro-climatic data

In order to represent the hydro-climate variability of the catchment, a 30-year period (1982–2011) was chosen according to data availability and quality. Precipitation and temperature data were interpolated based on respectively 12 and 8 stations (Fig. 1) using the inverse distance weighted method on a 5km x 5km grid. Since very few measurements were available outside the river valleys, elevation effects on precipitation and temperature distribution were considered using the SRTM digital elevation model (Fig. 1). In a previous study, Ruelland et al. (2014) examined the sensitivity of the GR4j hydrological model to different ways of interpolating climate forcing on this basin. Their results showed that a dataset based on a constant lapse rate of 6.5°C/km for temperature and no elevation effects for precipitation provided slightly better simulations of the discharge over the last 30 years.

However, since the current study also seeks to reproduce the seasonal dynamics of snow accumulation and melt, it was decided to rely on a mean monthly orographic gradient estimated from the precipitation observed series (Fig. 1). Potential evapotranspiration (PE) was computed using the following formula proposed by Oudin et al. (2005):

$$\mathrm{PE} = \frac{\mathrm{R_e}}{\lambda\rho} \times \frac{\mathrm{T} + K_2}{K_1} \quad \text{if} \;\; \mathrm{T} + K_2 > 0 \quad \text{else} \;\; \mathrm{PE} = 0 \tag{1}$$

where PE is the rate of potential evapotranspiration ($\mathrm{mm.d^{-1}}$), $\mathrm{R_e}$ is the extraterrestrial radiation ($\mathrm{MJ.m^{-2}.d^{-1}}$), λ is the latent heat flux (2.45 $\mathrm{MJ.kg^{-1}}$), ρ is the density of water ($\mathrm{kg.m^{-3}}$), T is the mean daily air temperature (°C) and K_1 and K_2 are fitted parameters (for more details on the values of K_1 and K_2, see Hublart *et al.* (2014)).. Water abstractions for irrigation were estimated using information on historical water allocations provided by the Chilean authorities. Because these abstractions are likely to influence the hydrological behavior of the catchment during recession and low-flow periods, they were added back to the gauged streamflow in Rivadavia before calibrating the models. In addition to streamflow data, remotely-sensed data from the MODerate resolution Imaging Spectroradiometer (MODIS) sensor were used to estimate the seasonal dynamics of snow accumulation and melt processes over a 9-year period (2003–2011). Daily snow cover products retrieved from NASA's Terra (MOD10A1) and Aqua (MYD10A1) satellites were combined into a single, composite 500-m resolution product to reduce the effect of swath gaps and cloud obscuration. The remaining data voids were subsequently filled using a linear temporal interpolation method.

2.3. Hydrological functioning of the catchment

2.3.1. Precipitation variability

Among the primary factors that control the hydrological functioning of the catchment is the high seasonality of precipitation patterns. Precipitation occurs mainly between June and August when the South Pacific High reaches its northernmost position. Most of the annual precipitation falls as snow at high elevations, where it accumulates in seasonal snow packs that are gradually released from October to April. The El Niño Southern Oscillation (ENSO) represents the largest source of climate variability at the interannual timescale (e.g. Montecinos and Aceituno, 2003). Anomalously wet (dry) years in the region are generally associated with warm (cold) El Niño (La Niña) episodes and a simultaneous weakening (strengthening) of the South Pacific High. It is worth noting, however, that some very wet years in the catchment can also coincide with neutral to weak La Niña conditions, as in 1984, while several years of below-normal precipitation may not exhibit clear La Niña characteristics (Verbist et al., 2010; Jourde et al., 2011). These anomalies may be due to other modes of climate variability affecting the Pacific basin on longer timescales. The Interdecadal Pacific Oscillation (IPO), in particular, has been shown to modulate the influence of ENSO-related events according to cycles of between 15 and 30 years (Quintana and Aceituno, 2012). Recent shifts in the IPO phase occurred in 1977 and 1998 and may be responsible for the highest frequency of humid years during the 1980s and the early 1990s when compared to the late 1990s and the 2000s.

2.3.2. Catchment-scale water balance and dominant processes

Notwithstanding this significant climate variability, a rough estimate of the catchment water balance can be given for the period 2003–2011 using the data presented in the previous subsection and additional information available in the literature. Spatially averaged precipitation ranges from a low of 80 mm in 2010 to an estimated high of 190 mm in 2008. Evapotranspiration from non-cultivated areas is sufficiently low to be reasonably neglected at the basin scale (Kalthoff et al., 2006). By contrast, water losses from the cultivated portions of the basin are likely to be around 10 $\mathrm{mm.yr^{-1}}$ (Hublart et al., 2014). At high elevations, sublimation plays a much greater role than evapotranspiration. Mean annual sublimation rates over two glaciers located in similar, neighbouring catchments have been estimated to be about 1 $\mathrm{mm.d^{-1}}$ (see e.g. MacDonell et al., 2013). Thus, a first estimate of the annual water loss associated with snow sublimation can be made by multiplying, for each day of the period, the proportion of the catchment covered with snow by an average rate of 1 $mm.d^{-1}$. This leads to a mean annual loss of 70 mm between 2003 and 2011. Note that this value is of the same order of magnitude as those obtained by Favier et al. (2009) using the Weather Research and Forecasting regional-scale climate model. Mean annual discharge per unit area varies from a minimum of 20 mm in 2010 to a maximum of 140 mm in 2003. Interestingly, runoff coefficients exceed 100% during several years of the period (in 2003, 2006, 2007 and 2009), indicating either an underestimation of precipitation at high elevations, as suggested by Favier et al. (2009), or a delayed contribution of groundwater to surface flow from one year to another (Jourde et al., 2011).

Groundwater movement in the catchment is mainly from the mountain blocks toward the valleys and then northward along the riverbed. In the mountains, groundwater flow and storage are controlled primarily by the presence of secondary permeability in the form of joints and fractures (Strauch et al., 2006). The unconfined valley-fill aquifers are replenished by mountain front recharge along the valley margins and by infiltration through the channel bed along the losing river reaches (Jourde et al., 2011). Their hydraulic conductivity and saturated thickness range from about 10 $m.d^{-1}$ and 40 m respectively in the upper part of the catchment to more than 30 $m.d^{-1}$ and 60 m respectively at the outlet (CAZALAC, 2006), allowing a rapid transfer of water to the hydraulically connected surface streams. Pourrier et al. (2014) studied flow processes and dynamics in the headwaters of the neighbouring Turbio River catchment; yet very little remains currently known about the emergent processes taking place at the catchment scale.

3. METHODS

3.1. Multiple-hypothesis modeling framework

In order to evaluate various numerical representations of the catchment functioning, a multiple-hypothesis modeling framework inspired by previous studies in literature was developed. All the models built within this framework are lumped hypotheses run at a daily time step. The modeling process was decomposed into three modules and six model-building decisions. Each module deals with a different aspect of the precipitation–runoff relationship through one or more decisions (Fig. 2): snow accumulation (A) and melt (B), runoff generation (C), redistribution (D) and delay (E) of water fluxes, and natural storage effects (F). Each of these decisions is provided with a set of alternative modeling options, which are named by concatenating the following elements: first a capital letter from A to F referring to the decision being addressed, then a number from 1 to 3 to distinguish between several competing architectures and, finally, a lower case letter from *a* to *c* to indicate different parameterizations of the same architecture. Model hypotheses are named by concatenating the names of the six modeling options used to build them (see Table 4). The models designed within this framework share the same overall structure (based on the same series of decisions) but differ in their specific formulations within each decision.

The model-building decisions can be divided into two broad categories. The first pertains to the production of fluxes from conceptual stores (decisions B, C and F). The second concerns the allocation and transmission of these fluxes using the typical junction elements and lag functions (decisions A, D and E) described in Fenicia *et al.* (2011). Junction elements can be defined as “zero-state” model components used to combine several fluxes into a single one (option D2) or split a single flux into two or more fluxes (options A1 and D3). Lag functions are used to reflect the travel time (delay) required to convey water from one conceptual store to another or from one or more conceptual stores to the basin outlet. They usually consist of convolution operators (option E2), although conceptual stores may also do the trick. Modeling options in which water fluxes are left unchanged are labelled as “No operation” options in Fig. 2. Water fluxes and state variables are named using generic names (from Q1 to Q6 and from S1 to S4, respectively) to ensure a perfect modularity of the framework. Further details on the alternative options provided for each decision are given in the following subsections. Note that some combinations of modeling options were clearly incompatible with one another (options C1 and C2, for instance, cannot work with option D2). As a result, these combinations were removed from the framework.

Another important feature of this modular framework is the systematic smoothing of all model thresholds using infinitely differentiable approximants, as recommended by Kavetski and Kuczera (2007) and Fenicia *et al.* (2011). The purpose here is twofold: first, to facilitate the calibration process by removing any unnecessary (and potentially detrimental) discontinuities from the gradients of the objective functions; and second, to provide a more realistic description of hydrological processes across the catchment (Moore, 2007).

3.1.1.Snow accumulation and melt (decisions A and B)

Snow accumulation and melt components deal with the representation of snow processes at the catchment scale. All modeling options rely on a single conceptual store to accumulate snow during the winter months and release water during the melt season. Decision A refers to the partitioning of precipitation into rain, snow or a mixture of rain and snow. Decision B refers to the representation of snowmelt processes. Option A1 is the only hypothesis implemented to evaluate the relative abundance of rain and snow. A logistic distribution is used in this option instead of usual temperature thresholds to implicitly account for spatial variations in rain/snow partitioning over the catchment. In contrast, three modeling options drawing upon the temperature-index approach (Hock, 2003) are available for the evaluation of snowmelt rates (options B1a, B1b, B1c). Option B1a relies on a constant melt factor while options B1b and B1c allow for temporal variability in the melt factor to reflect seasonal changes in the energy available for melt. A recent example of option B1c can be found in Clark *et al.* (2009). Option B1b has been previously applied by Schreider *et al.* (1997) but at the grid cell scale. Finally, it is worth noting that a smoothing kernel proposed by (Kavetski and Kuczera, 2007) was introduced in the state equation of the snow reservoir to ignore residual snow remaining in the reservoir outside the snowmelt season.

3.1.2. Runoff generation (decision C)

Runoff generation components determine how much of a rainfall or snowmelt event is available for runoff, lost through evapotranspiration or temporarily stored in soils and surface depressions. Many models rely on a conceptual store to keep track of the catchment moisture status and generate runoff as a function of both current and antecedent precipitation. Here, an assortment of four commonly used methods is available. Option C1 is the only one in which no moisture accounting store is required to estimate the contributing rainfall or snowmelt (see Fig. 3). Actual evapotranspiration then represents the only process involved in the production of runoff from precipitation or snowmelt. The remaining options make use of moisture accounting stores and distribution functions (see Table 1) to estimate the proportion of the basin generating runoff. An important distinction is made between option C2, in which runoff generation occurs only during rainfall or snowmelt events, and option C3, in which a leakage from the moisture accounting store remains possible even after rainfall or snowmelt has ceased. Examples of these two moisture accounting options can be found, respectively, in the HBV (e.g. Seibert and Vis, 2012) and PDM (Moore, 2007) rainfall-runoff models. Alternative distribution functions are available in the literature, for instance in the GR4j (Perrin *et al.*, 2003) and FLEX (Fenicia *et al.*, 2008b) models, but the rationale behind their use remains the same. Actual evapotranspiration is computed from the estimated PE using either a constant coefficient (option C1) or a function of the catchment moisture status (options C2 and C3).

3.1.3. Runoff transformation and routing (decisions D to F)

Runoff transformation components account for all the retention and translation processes occurring as water moves through the catchment. In practice, junction elements (decision D) and lag functions (decision E) are typically combined with one or more conceptual stores (decision F) to represent the effects of different flow pathways on the runoff process (both timing and volume). Additional elements in the form of lag functions or conceptual stores can also be used to reflect water routing in the channel network. However, in this study channel routing elements were considered useless at a daily time step. All the modeling options available for decision F consist of two stores. These can be arranged in parallel (options F1a and F1b), in series (options F2a and F2b), or in a combination of both (options F3a and F3b). In each case, one of the stores has a nonlinear behavior while the other reacts linearly. Two types of nonlinear response are provided: one that relies on smoothed thresholds and different storage coefficients (options F1b, F2b and F3b), and the other that relies on power laws (options F1a, F2a and F3a). Options F1a and F1b are based on the classical parallel transfer function used in many conceptual models, such as the PDM (Moore, 2007) and IHACRES (Jakeman *et al.*, 1993) models, where one store stands for a relatively quick catchment response and the other for a slower response. The structure of options F3a and F3b is very close to the response routine of the HBV model (e.g. Seibert and Vis, 2012). Note that some combinations of modeling options were deemed unacceptable and thus not considered (e.g. D3–E1–F1a or D3–E1–F1b).

3.2. Multi-objective optimization

3.2.1. Principle

In optimization problems with at least two conflicting objectives, a set of solutions rather than a unique one exists because of the trade-offs between these objectives. A Pareto-optimal solution is achieved when it cannot be improved upon without degrading at least one of its objective criteria. The set of Pareto-optimal solutions for a given model is often called the "Pareto set" and the set of criteria corresponding to this Pareto set is usually referred to as the "Pareto front".

3.2.2. The NSGA–II algorithm

The Non-dominated Sorted Genetic Algorithm II (NSGA–II) (Deb, 2002) was selected to calibrate the models implemented within the multiple-hypothesis framework. This algorithm has been used successfully in a number of recent hydrological studies (see e.g. Khu and Madsen, 2005; Bekele and Nicklow, 2007; De Vos and Rientjes, 2007; Fenicia et al., 2008a; Shafii and De Smedt, 2009) and has the advantage of not needing any additional parameter (other than those common to all genetic algorithms, i.e. the initial population and the number of generations). Its most distinctive features are the use of a binary tournament selection, a simulated binary crossover and a polynomial mutation operator. For brevity's sake, the detailed instructions of the algorithm and the conditions of its application to rainfall-runoff modeling cannot be discussed further here. Instead, the reader is referred to the aforementioned literature.

3.2.3. Simulation periods and assessment criteria

The simulation period was divided into a rather dry calibration period (1997–2011) and a relatively humid validation period (1982–1996). These two periods were chosen based on data availability to represent contrasted climate conditions: the two periods are separated by a shift in the IPO index, as explained in Sect 2.3.1.

Four criteria were chosen to evaluate the models built within the multiple-hypothesis framework. The first three of them are common to both calibration and validation periods while the fourth criterion differs between the two.

The first criterion (NSE) is the related to the estimation of high flows and draws upon the Nash-Sutcliffe Efficiency metric:

$$\mathrm{Crit1} = 1 - \mathrm{NSE} = \sum_{d=1}^{N} \left(Q_{obs}^{d} - Q_{sim}^{d}\right)^2 \Big/ \sum_{d=1}^{N} \left(Q_{obs}^{d} - \overline{Q_{obs}}\right)^2 \qquad (2)$$

Where Q_{obs}^{d} and Q_{sim}^{d} are the observed and simulated discharges for day d, and N is the number of days with available observations.

The second criterion (NSE_{log}) is related to the estimation of low flows and draws upon a modified, log version of the first criterion:

$$\mathrm{Crit2} = 1 - \mathrm{NSE}_{\mathrm{log}} = \sum_{d=1}^{N} \left(\log\left(Q_{obs}^{d}\right) - \log\left(Q_{sim}^{d}\right)\right)^{2} \Big/ \sum_{d=1}^{N} \left(\log\left(Q_{obs}^{d}\right) - \log\left(\overline{Q_{obs}}\right)\right)^{2} \quad (3)$$

The third criterion quantifies the mean annual volume error (VE_M) made in the estimation of the water balance of the catchment:

$$\mathrm{Crit3} = \mathrm{VE_M} = \sum_{y=1}^{N_{years}} \left(\left|V_{obs}^{y} - V_{sim}^{y}\right| / V_{obs}^{y}\right) / N_{years} \quad (4)$$

Where V_{obs}^{y} and V_{sim}^{y} are the observed and simulated volumes for year y, and N_{years} is the number of years of the simulation period.

The fourth criterion (Crit4) differs between the two simulation periods. In calibration, snow-covered areas (SCA) estimated from the MODIS data were used to evaluate the consistency of snow-accounting modeling options in terms of snow presence or absence at the catchment scale. The objective was to quantify the error made in simulating the seasonal dynamics of snow accumulation, storage and melt processes. Following Parajka and Blöschl (2008), the snow error (SE) was defined as the total number of days when the snow-accounting store of options B1a, B1b and B1c disagreed with the MODIS data as to whether snow was present in the basin (Fig. 4). The number of days with simulation errors is eventually divided by the total number of days with available MODIS data to express SE as a percentage.

In validation, a cumulated volume error was used to replace the snow error criterion that could not be computed due to a lack of remotely-sensed data over this period:

$$\mathrm{Crit4} = \mathrm{VE_C} = \left|\sum_{y=1}^{N_{years}} V_{obs}^{y} - \sum_{y=1}^{N_{years}} V_{sim}^{y}\right| \Big/ \sum_{y=1}^{N_{years}} V_{obs}^{y} \quad (5)$$

3.3. Model selection, model analysis and ensemble modeling

Finally, a total of 72 model structures were implemented and tested within the multi-objective and multiple-hypothesis frameworks. In addition to their names and for purposes of simplicity, these 72 model hypotheses are given a number from 1 to 72 corresponding to their order of appearance in the simulation process (see e.g. Sect 4.1.).

Model hypotheses can be thought of as points x in the space of performance measures. One possible way to locate these points in space is to consider that each coordinate $(x_i)_{i=1\ldots4}$ of x is given by the best performance obtained along the Pareto front of model x with respect to the i^{th} criterion described in Sect 3.3.2. A clustering technique based on the fuzzy c-means algorithm (Bezdek et al., 1983) and the initialization procedure developed by Chiu (1994) was chosen to explore this multi-objective space and identify natural groupings among model hypotheses. To facilitate comparison between calibration and validation, the clustering operations were repeated independently for each period. The whole experiment, from model building to multi-objective optimization and cluster identification, was repeated several times to ensure that the final composition of the clusters remains the same.

Once the composition of each cluster was established, it was possible to identify a set of 'best-performing' clusters for each simulation period, i.e. a set of clusters with the smallest Euclidian distances to the origin of the objective space. The model structures of these 'best-performing' clusters can be regarded as equally acceptable representations of the system. An important indicator of structural uncertainty is the extent to which the simulation bounds derived from the Pareto sets of these models reproduce the various features of the observed hydrograph. The overall uncertainty envelope should be wide enough to include a large proportion of the observed discharge but not so wide that its representation of the various aspects of the hydrograph (rising limb, peak discharge, falling limb, baseflow) becomes meaningless. In this study, priority was given to maintaining at its lowest value the number of outlying observations before searching for the best combination of models which minimized the envelope area. This was achieved iteratively through the following steps:

1. Start with an initial ensemble composed of the N_{max} models identified as members of the best-performing clusters in both calibration and validation (i.e. models which fail the validation test are ruled out).
2. From now on, consider only the calibration period.
   Add up the N_{max} individual simulation envelopes that can be obtained from the Pareto sets of the N_{max} models (hereafter referred to as the 'Pareto-envelopes').
3. Estimate the maximum number of observations enclosed within the resulting overall envelope, $N_{\text{obs}}(N_{\text{max}})$, and calculate the area of this envelope, $Area(N_{\text{max}})$.
4. For $k = 1$ to N_{max}
   a. Identify the $\binom{N_{max}}{N_{max} - k}$ possible combinations of N_{max} models taken $N_{max} - k$ at a time.
   b. For each of these combinations
      - Add up the individual Pareto-envelopes of the $N_{max} - k$ models and calculate the number of observations enclosed within the bounds of the resulting overall envelope, $N_{\text{obs}}(N_{\text{max}} - k)$.
      - If $N_{\text{obs}}(N_{\text{max}} - k) = N_{\text{obs}}(N_{\text{max}})$
        If $Area(N_{\text{max}} - k) < Area(N_{\text{max}} - k + 1)$
        Accept the current combination.
        If $N_{\text{obs}}(N_{\text{max}} - k) < N_{\text{obs}}(N_{\text{max}})$
        Reject the current combination.
   c. If all the possible combinations of $N_{\text{max}} - k$ models are rejected, break the loop. The final ensemble of models to consider is the last accepted combination of $N_{\text{max}} - k + 1$ models.

4. RESULTS

4.1. Model hypotheses evaluation

4.1.1. Cluster analysis

The 72 model hypotheses can be grouped into 5 clusters in calibration and 6 in validation. Table 3 displays the coordinates of the cluster centroids and gives, for each cluster, the number of points with membership values above 50%. Figure 5 shows the projections of these clusters onto three possible two-dimensional (2D) subspaces of the objective space (the three other subspaces being omitted for brevity's sake). Each cluster is given a rank (from 1 to 5 or 6) reflecting its distance from the origin of the coordinate system. As is evident from both Fig. 5 and Table 3, most of the best-performing structures can be found in Cluster 1. This is particularly clear in the planes defined by the high-flow (Crit1) and low-flow (Crit2) criteria (Figure 5), where all clusters tend to line up along a diagonal axis (dashed line). In contrast, a small trade-off between Cluster 1 and Cluster 2 can be observed in calibration in the plane defined by the high-flow (Crit1) and volume error (Crit3) criteria: models from Cluster 2 (respectively Cluster 1) tend to perform slightly better than those from Cluster 1 (respectively Cluster 2) with respect to Crit3 (respectively Crit1). However, this trade-off disappears in validation. Similar comments can be made about the other 2D subspaces (not shown here). In the following analysis, Cluster 1 will be considered as the only best-performing cluster. This cluster encompasses 24 members in calibration as against 15 in validation, indicating that several model structures do not pass the validation test (namely models no. 30, 32, 49, 52, 53, 55, 66, 67, 69 and 72, as shown in Table 4).

Several observations can be made regarding the composition of Cluster 1 in both simulation periods. As can be seen from the values listed in Table 4, it is not possible to pick out a single, unambiguous model hypothesis that would perform better than the others with respect to all criteria.

On the one hand, there appears to be several equally acceptable structures for each individual criterion. Models no. 22 (A1–B1a–C3–D2–E1–F2b), 46 (A1–B1b–C3–D2–E1–F2b) and 54 (A1–B1c–C1–D3–E2–F1b), for instance, yield very similar values of the high-flow criterion (Crit1), despite some differences in their modeling options. This illustrates the equifinality of model structures in reproducing one aspect of the system behavior. On the other hand, some structures seem more appropriate to the simulation of high flows or snow dynamics while others appear to be better at reproducing low flows or estimating the annual water balance of the catchment. This indicates trade-offs between model structures in reproducing several aspects of the system behavior. It is however possible to identify some recurring patterns among the modeling options present in (or absent from) Cluster 1 in both periods. First, option B1c is the most represented snowmelt-accounting hypothesis, despite an increase in the number of alternative options (B1a, B1b) in validation. More strikingly, option C2 is totally absent from Cluster 1 in both periods. Single-flux combinations (C1–D1 and C3–D2) and their splitting counterparts (C1–D3 and C3–D1) tend to be equally well-represented, thus providing evidence of significant equifinality among these conceptual representations. Finally, runoff transformation options based on a threshold-like behavior (F1b, F2b and F3b) account for 75% of model hypotheses in calibration and over 90% in validation. In particular, option F3a turns out to be completely absent from Cluster 1 in both periods while models based on option F2a (no. 49, 55, 67 and 69) fail the validation test. On the opposite, option F2b is particularly well-represented.

4.1.2. Pareto analysis

In general, valuable insight can be gained from the mapping of Pareto fronts in the space of performance measures. While a full description of all the Pareto fronts obtained in calibration is not possible here due to space limitations, two model hypotheses are used to illustrate this point. Figure 6 shows the Pareto-optimal solutions of models no. 49 (A1–B1c–C1–D1–E1–F2a) and 50 (A1–B1c–C1–D1–E1–F2b) plotted in two dimensions for different combinations of two of the four objective functions used in calibration. Note that these two models differ only in their runoff transformation options (F2a *vs.* F2b) so that the comparison can be made in a controlled way. Trade-offs between the high-flow (Crit1) and low-flow (Crit2) criteria are clearly more important with option F2a (Fig. 6a) than with option F2b (Fig. 6b). This means that option F2a is less efficient in reproducing simultaneously high and low flows and explains why this option disappears from Cluster 1 in validation. By contrast, the other pairs of criteria (Crit1–Crit3, Crit1–Crit4) displayed in Fig. 6 appear to be less useful in differentiating between the two models.

Further insight into the structural strengths and weaknesses of model hypotheses can be obtained by determining how parameter values vary along the Pareto fronts of the models. A large 'Pareto range' in some parameters indicates structural deficiencies in the corresponding model components (see e.g. Gupta *et al.*, 1998) or a lower sensitivity of model outputs to those parameters (Engeland *et al.*, 2006). For purposes of clarity, Fig. 7 focuses on eight illustrative structures identified as members of Custer 1 in calibration. The models are paired in such a way that two models of the same pair differ in only one modeling option. Thus, the effects of potential interactions between model constituents are more likely to be detected. Parameter values are normalized using the lower and upper limits given in Table 2 so that all of them lie between 0 and 1. Different colors are used to indicate the parameter sets associated with the smallest high-flow (in black), low-flow (in red), volume (in blue) and snow (in green) errors. To what extent these colored solutions converge toward the same parameter values or diverge from each other determines the level of parameter identifiability of each model hypothesis. As regards snow-accounting options, a distinction can be made between snow accumulation paramaters (T_S and m_S), whose ranges of variation appear to be large in all cases, and snowmelt parameters (T_M, f_M, r_1, r_2, f_1, f_2), whose levels of identifiability depend on interactions with the other model components. In Fig. 7a, the Pareto range of snowmelt parameters decreases in width when moving from option B1a to B1b and using the combination of options C3–D2–E1. Yet changing this combination into C3–D1–E2 has the opposite effect (Fig. 7b): parameter uncertainty now decreases when moving from option B1b to B1a. As regards runoff transformation parameters (α, N_b, K_2, K_3, δ, S_C and K_4), the black and red solutions are closer to each other when options F2b (Fig. 7a, 7b and 7c) and F1b (Fig. 7d) are used. By contrast, options F2a (Fig. 7c) and F1a (Fig. 7d) require very different parameter sets to adequately simulate both low and high flows. Again, this suggests that runoff transformation options based on a threshold-like behavior may be more consistent with the observed data than those based on a power law relationship. It should be noted, however, that relatively large Pareto ranges in some runoff transformation parameters (e.g. K_2 and K_3) may still be required to obtain small volume and snow errors at the same time as high low-flow and high-flow performances (e.g. models no. 44 and 54). Interestingly, the black, red and blue solutions of models no. 49, 50, 53 and 54 also converge towards the same low values of parameter K_C (evapotranspiration coefficient) independently of runoff transformation options.

Drawing any conclusion at this stage about the links between parameter identifiability and model performance might be somewhat hazardous. Other examples (not shown here) show that a model structure may have highly identifiable parameter values in calibration and yet not be suited to the conditions prevailing in validation. Also, a reduction of parameter uncertainty as is the case with options F2b and F1b often comes with a greater number of parameters.

Finally, a better understanding of the reasons why some models, or modeling options, work better than others is provided by the simulation bounds (or Pareto-envelopes) derived from the Pareto sets of these models. Figure 8 shows the Pareto-envelopes of the SWE internal state variable obtained with three competing model hypotheses (no. 6, 30 and 54) differing only in their snowmelt-accounting options (respectively B1a, B1b and B1c). Note that only the last two of these models (30, 54) belong to Cluster 1 in calibration (see Table 4). Simulated snow accumulation starts later than expected with all modeling options (B1a, B1b and B1c). As will be further discussed in Sect 5.2., this is likely to indicate systematic errors in the input precipitation and/or MODIS-based SCA data. On the whole, the envelope widths suggest a reduction in the uncertainty associated with the prediction of snow seasonal dynamics when moving from option B1a to option B1c. This is consistent with the mean annual snow errors reported in Table 4, which are significantly lower with option B1c independently of the other model options. It must be acknowledged, however, that even this option (B1c) fails to capture the seasonal dynamics of snow accumulation and melt during several years of the period. The release of water from the snow-accounting store of model no. 54 continues well after the end of the observed snowmelt season in 2008, 2009, 2010 and 2011. On the contrary, the simulated snowmelt season tends to end sooner than expected with model no. 30 in 2003, 2004, 2005 and 2006. In that case, options B1b and B1c appear to be somewhat complementary.

4.2. Representation of structural uncertainties

This Section deals with the identification and use of an ensemble of equally acceptable model structures to quantify and represent the uncertainty arising from the system non-identifiability. Figure 9 shows the overall uncertainty envelope obtained with the 8 model structures whose combination minimizes the envelope area in calibration while holding constant the number of outlying observations (see Sect 3.3.). Over 82% of discharge observations are captured by the envelope in both simulation periods. Interestingly, this number exceeds the best N_{par} value obtained in calibration with the individual Pareto-envelopes (see Table 4), which shows how necessary it is to consider an ensemble of model structures. In validation, however, a better combination could be identified since several models of Cluster 1 display significantly higher N_{par} values (Table 4). On the whole, the comparison of the observed hydrograph with the simulation bounds of the envelope shows a good match of rising limbs and peak discharges in both simulation periods, but a less accurate fit of falling limbs during at least one major (in 1987–88) and two minor (in 2005–06 and 2007–08) events. The slower recession of the observed hydrograph might indicate a delayed contribution of one or more catchment compartments that cannot be described by any of the modeling options available in the multiple-hypothesis framework.

5. DISCUSSION & CONCLUSION

This study aimed at reducing structural uncertainty in the modeling of a semi-arid Andean catchment where lumped conceptual models remain largely under-used. To overcome the current lack of information on model adequacy in this catchment, a modular modeling framework (MMF) relying on six model-building decisions was developed to generate 72 competing model structures. Four assessment criteria were then chosen to calibrate and evaluate these models over a 30-year period using the concept of Pareto-optimality. This strategy was designed to characterize both the parameter uncertainty arising from each model's structural deficiencies (i.e. model inadequacy) and the ambiguity associated with the choice of model components (i.e. model non-uniqueness). Finally, a clustering approach was taken to identify natural groupings in the multi-objective space. Overall, the greatest source of uncertainty was found in the connection between runoff generation and runoff transformation components (decisions D and E). However, the results also showed a significant drop in the number of plausible representations of the system. After validation, 14 model structures among the 24 identified in calibration as the best-performing ones were finally considered as equally acceptable.

Interestingly, both rejected and accepted hypotheses appeared closely related to particular types of snowmelt-accounting (decision B), runoff generation (decision C) and runoff transformation (decision D) modeling options, suggesting possible links to some physical features of the catchment. For instance, the frequent occurrence of option C1 and the absence of option C2 among the set of best-performing structures indicate that moisture-accounting components may not be essential to the conceptual modeling of this catchment. Most of the land cover is, indeed, dominated by barren to sparsely vegetated exposed rocks, boulders and rubble with poor soil development outside the valleys. This setting may also explain the relatively low values of parameter K_C obtained with the black, red and blue solutions shown in Fig. 6. Likewise, the frequency of options F2a and F2b in the best-performing cluster suggests that the catchment actually behaves as a 'serial' system. The overall organization of fluxes in the catchment, from high elevations toward the valleys and then northward to the outlet, can be conceptualized as a series of two hydraulically connected reservoirs: one standing for the granitic mountain blocks (upstream reservoir) and the other for the alluvial valleys (downstream reservoir). Similar results were also obtained for smaller catchments in Luxembourg characterized by relatively impervious bedrocks and lateral water flows (Fenicia et al., 2014). The results also provided some evidence of a strong threshold behavior at the catchment scale (options F1b, F2b and F3b) compared to the smoother power laws of options F1a, F2a and F3a. However, further research would be needed to track the origin of this behavior, which might be related at some point to connectivity levels in the fractured and till-mantled areas of the mountain blocks. As regards snowmelt, the frequent occurrence of option B1c in the best-performing cluster in calibration may indicate a need to account for processes which the degree-day method implemented in option B1a does not fully capture. In semi-arid central Andes (29–30°S), small zenith angles and a thin, dry and cloud-free atmosphere during most of the year make incoming shortwave radiation the most important source of seasonal variations in the energy available for melt (e.g. Pellicciotti et al., 2008; Abermann et al., 2013). While this dominant source of energy cannot be accounted for by temperature alone, the seasonal timing of snowmelt is also expected to show a greater year-to-year stability, which may explain the relative success of option B1c when compared to option B1b. Of course, these hypothesized relationships between some physical characteristics of the catchment and specific modeling options need to be further qualified. Differentiating between physically adequate and purely numerical solutions will always seem somewhat hazardous in the case of lumped conceptual models. For instance, a small number of models among those identified as the best-performing ones also rely on parallel (F1a, F1b) and intermediate (F3b) runoff transformation options. Also, the relative proportions of snowmelt-accounting options B1a, B1b and B1c, appears much more balanced in validation, where no snow error criterion could be applied, than in calibration. Although this was not our objective in this paper, comparative studies including several similar or contrasted catchments would be required to better understand how different model structures relate to different physical settings. Such understanding is of primary importance to the choice of conceptual models in climate change impact studies.

Another important issue related to model identification is the extent to which the 'principle of parsimony' can be applied to differentiate between a large number of model hypotheses. Many authors rightly consider that a maximum of 5 to 6 parameters should be accepted in calibration when using a single objective function. Efstratiadis and Koutsoyiannis (2010) extended this empirical rule to the case of multi-objective schemes by allowing « a ratio of about 1:5 to 1:6 between the number of criteria and the number of parameters to optimize ». For a multi-objective scheme based on four criteria (as in the present study), this leads to consider 20 to 24-parameter models as still being parsimonious. This will certainly seem unreasonable to many modelers because, as Efstratiadis and Koutsoyiannis (2010) also pointed out, the various criteria used are generally not independent of each other. In our case, for instance, the information added by the low-flow criterion may not be so different from that already introduced by the high-flow criterion. By contrast, the snow criterion tends to add new information on the snow-related parameters. From this perspective, it is noteworthy that most rejected hypotheses among the 24 identified in calibration as members of Cluster 1 had more than 11 free parameters, with only one having 9 parameters. The principle of parsimony, however, cannot be used to further discriminate between the remaining 14 best-performing hypotheses. For instance, model no. 54 (12 parameters) performs better than model no. 2 (9 parameters) with respect to the high-flow criterion.

Eventually, the number of models used to represent structural uncertainty was reduced by searching for which minimal set of models maximized the number of observations covered by the ensemble of Pareto-envelopes. It is important to make clear that model inadequacy and non-uniqueness were evaluated here in non-probabilistic terms. In particular, the Pareto-envelopes derived for each model structure quantify only the uncertainty arising from the trade-offs between competing criteria and do not have a predefined statistical meaning (Engeland et al., 2006). Consequently, the overall simulation bounds shown in Figure 8 cannot be easily interpreted as 'confidence bands'. Although discussing the adequacy of non-probabilistic approaches to structural uncertainty was far beyond the scope of this study, it is interesting to analyze the reasons why between 15% and 20% of the observations remained outside the overall simulated envelope in both calibration and validation. To a large extent, this lack of performance can be attributed either to an insufficient coverage of the hypothesis and objective spaces or to uncertainties in the precipitation and streamflow data that were overlooked in this study.

First, the choice of Pareto-optimality to characterize structural uncertainty can be criticized for leading to the rejection of many behavioral parameter sets (i.e. being close to, but not part of, the Pareto front) that might have been Pareto-optimal with different performance measures, calibration data or input errors (e.g. Freer et al., 2003; Beven, 2006). Also, this concept should not be confused with that of equifinality. Both notions agree that it is not possible to identify a single, best solution to the calibration problem and that multiple parameters sets should be retained to give a proper account of model uncertainty. However, the Pareto set of solutions represents the minimum parameter uncertainty that can be achieved when several criteria are considered simultaneously with no *a priori* preference for one over the others (Gupta et al., 2003). By contrast, two parameter sets are said to be equifinal (in a statistical sense) if they can be regarded as equally acceptable with respect to a given model outcome. For a proper assessment of parameter equifinality, more probabilistic approaches should be taken (Madsen, 2000; Huisman *et al.*, 2010). In the context of multiple-hypothesis testing, a meticulous selection of the assessment criteria is also critical to avoid rejecting some modeling options for the wrong reasons. For instance, the snow error criterion was shown to have a great influence on the identification of snow-accounting components, as much more ambiguity between the various available options was observed during the validation period when this criterion could not be used. Also, like any other multiple-hypothesis framework, the MMF developed in this study suffers from an insufficient coverage of the hypothesis space (Gupta et al., 2012). The parameterization of evapotranspiration, for example, was not considered as an independent model-building decision. Only one formula was applied to calculate potential evapotranspiration and the possibility to retrieve actual evapotranspiration from downstream water stores was not provided. Likewise, the runoff transformation process was described using only two water stores, of which only one was assumed to have a nonlinear behavior. Future work to improve the conceptual modeling of the Claro River catchment should include the testing of new or refined hypotheses to allow for the use of additional auxiliary data (e.g. observed snow heights, irrigation water-use).

More fundamentally, our ability to discriminate among the competing model hypotheses was constrained by inevitable errors in the input and output data sets. In particular, the comparison of simulated SWE levels and MODIS-based SCA estimates revealed some uncertainty in the estimation of precipitation inputs and confirmed previous results obtained by Favier et al. (2009). Some precipitation events occurring in the early winter may not be captured by the gauging network (< 3200 m a.s.l.) used for the interpolation of precipitation across the catchment. These errors may add to systematic volume errors caused by wind, wetting and evaporation losses at the gauge level, leading to an overall underestimation of precipitation, as indicated by the rough estimate of the catchment-scale water balance given in Sect 2. It was also possible to highlight some errors in the streamflow data. The observed streamflow was 'naturalized' by simply adding back the estimated historical water abstractions (Sect. 2.2). When applied on a daily basis, this process inevitably adds some uncertainty to streamflow values because a significant part of surface-water abstractions actually return to the river system within a few days due to conveyance and field losses. In general, ignoring these return flows would lead to overestimating daily natural flows. In this paper, however, the actual water withdrawals were not known with precision but only as percentages of the nominal water rights – these percentages being fixed on a monthly basis by the authorities to account for variations in water availability. The combined impact of streamflow and precipitation errors on the assessment of structural uncertainty thus remained unknown. Further research is currently underway to integrate the effects of water abstractions and crop water-use in the hydrological modeling process (Hublart et al., 2015; see also Kiptala et al., 2014 for another approach). From a multiple-hypothesis perspective, the modeling of irrigation water-use should be regarded as a testable model component in its own right.

Acknowledgements The authors are very grateful to the Centro de Estudios Avanzados en Zonas Áridas (CEAZA) for its essential logistic support during the field missions and to Gustavo Freixas from the *Dirección General de Agua* (Chile) for providing the necessary streamflow data. The authors also thank S. Lhermitte, D. López and S. MacDonell for providing the MODIS data used in this study and S. Gascoin for informal advice and much useful discussion. Moreover, the authors thank the two anonymous reviewers for their interest to this work and for their useful comments that helped to improve the article.

REFERENCES

Abermann, J., Kinnard, C., and MacDonell, S.: Albedo variations and the impact of clouds on glaciers in the Chilean semi-arid Andes, J. Glaciol., 60, 183–191, 2013.

Bekele, E. G. and Nicklow, J. W.: Multi-objective automatic calibration of SWAT using NSGA-II, J. Hydrol., 341, 165–176, 2007.

Beven, K.: Prophecy, reality and uncertainty in distributed hydrological modelling, Adv. Water Resour., 16, 41–51, 1993.

Beven, K.: A Manifesto for the Equifinality Thesis, J. Hydrol., 320, 18–36, 2006.

Bezdek, J. C., Ehrlich, R., and Full, W.: FCM: The fuzzy c-means clustering algorithm, Comput. Geosci., 10, 191–203, 1983.

Birkel, C., Tetzlaff, D., Dunn, S. M., and Soulsby, C.: Towards a simple dynamic process conceptualization in rainfall–runoff models using multi-criteria calibration and tracers in temperate, upland catchments. Hydrol. Process., 24, 260–275, doi: 10.1002/hyp.7478, 2010.

Blöschl, G. and A. Montanari: Climate change impacts–throwing the dice?, Hydrol. Process., 24, 374–381, 2010.

Boyle, D. P., Gupta, H. V., and Sorooshian, S.: Toward improved calibration of hydrologic models: Combining the strengths of manual and automatic methods, Water Resour. Res., 36, 3663–3674, doi:10.1029/2000WR900207, 2000.

Buytaert, W. and K. Beven: Models as multiple working hypotheses: hydrological simulation of tropical alpine wetlands, Hydrol. Process., 25, 1784–1799, 2011.

Capell, R., Tetzlaff, D., and Soulsby, C.: Can time domain and source area tracers reduce uncertainty in rainfall-runoff models in larger heterogeneous catchments?, Water Resour. Res., 48, W09544, doi:10.1029/2011WR011543, 2012.

Caviedes, C. N. and Paskoff, R.: Quaternary glaciations in the Andes of north-central Chile, J. Glaciol., 14, 155–169, 1975.

Centro del Agua para Zonas Áridas y semiáridas de América Latina y el Caribe (CAZALAC): Aplicación de metodologías para determinar la eficiencia de uso del agua – Estudio de caso en la Región de Coquimbo. Informe Técnico, Gobierno Regional, Santiago (Chile), 2006.

Chiu, S.: Fuzzy model identification based on cluster estimation, J. Intell. Fuzzy Syst., 2, 267– 278, 1994.

Clark, M. P., Slater, A. G., Barrett, A. P., Hay, L. E., McCabe, G. J., Rajagopalan, B., and Leavesley, G. H.: Assimilation of snow covered area information into hydrologic and landsurface models, Adv. Water Resour., 29, 1209–1221, 2006.

Clark, M. P., Slater, A. G., Rupp, D. E., Woods, R. A., Vrugt, J. A., Gupta, H. V., Wagener, T., and Hay, L. E.: Framework for Understanding Structural Errors (FUSE): A modular framework to diagnose differences between hydrological models, Water Resour. Res., 44, W00B02, doi:10.1029/2007WR006735, 2008.

Clark, M., Hreinsson, E. O., Martinez, G., Tait, A., Slater, A., Hendrikx, J., Owens, I., Gupta, H., Schmidt, J., and Woods, R.: Simulations of seasonal snow for the South Island, New Zealand, J. Hydrol., 48, 41–58, 2009.

Clark, M. P., Kavetski, D., and Fenicia, F.: Pursuing the method of multiple working hypotheses for hydrological modeling, Water Resour. Res., 47, W09301, doi:10.1029/2010WR009827, 2011.

Collet, L., Ruelland, D., Borrell-Estupina, V., Dezetter, A., and Servat, E.: Integrated modelling to assess long-term water supply capacity of a meso-scale Mediterranean catchment, Sci. Total Environ., 461–462, 528–540, 2013.

Coxon, G., Freer, J., Wagener, T., Odoni, N. A., and Clark, M. P.: Diagnostic evaluation of multiple hypotheses of hydrological behaviour in a limits-of-acceptability framework for 24 UK catchments, Hydrol. Process., doi:10.1002/hyp.10096, online first, 2013.

De Vos, N. J., and Rientjes, T. H. M.: Multi-objective performance comparison of an artificial neural network and a conceptual rainfall-runoff model, Hydrolog. Sci. J., 52, 397–413, doi: 10.1623/hysj.52.3.397, 2007.

Deb, K., Pratap. A, Agarwal, S., and Meyarivan, T.: A fast and elitist multiobjective genetic algorithm: NSGA-II, IEEE T. Evolut. Comput., 6, 181–197, 2002.

Dooge, J.: Looking for hydrologic laws, Water Resour. Res., 22, 46S–58S, doi:10.1029/WR022i09Sp0046S, 1986.

Dooge, J.: Searching for Simplicity in Hydrology, Surv. Geophys., 18, 511–534, 1997.

Efstratiadis, A., and Koutsoyiannis, D.: One decade of multi-objective calibration approaches in hydrological modelling: a review, Hydrolog. Sci. J., 55, 58–78, 2010.

Ehret, U., Gupta, H. V., Sivapalan, M., Weijs, S. V., Schymanski, S. J., Blöschl, G., Gelfan, A. N., Harman, C., Kleidon, A., Bogaard, T. A., Wang, D., Wagener, T., Scherer, U., Zehe, E., Bierkens, M. F. P., Di Baldassarre, G., Parajka, J., van Beek, L. P. H., van Griensven, A., Westhoff, M. C., and Winsemius, H. C.: Advancing catchment hydrology to deal with predictions under change, Hydrol. Earth Syst. Sci., 18, 649–671, doi:10.5194/hess-18-649-2014, 2014.

Engeland, K., Braud, I., Gottschalk, L., and Leblois, E.: Multi-objective regional modelling, J. Hydrol., 327, 339–351, 2006.

Favier, V., Falvey, M., Rabatel, A., Praderio, E., and López, D.: Interpreting discrepancies between discharge and precipitation in high-altitude area of Chile's Norte Chico region (26–32°S), Water Resour. Res., 45, W02424, doi:10.1029/2008WR006802, 2009.

Fenicia, F., McDonnell, J. J., and Savenije, H. H. G.: Learning from model improvement: On the contribution of complementary data to process understanding, Water Resour. Res., 44, W06419, doi:10.1029/2007WR006386, 2008a.

Fenicia, F., Savenije, H. H. G., Matgen, P., and Pfister, L.: Understanding catchment behavior through stepwise model concept improvement, Water Resour. Res., 44, W01402, doi:10.1029/2006WR005563, 2008b.

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

Fenicia, F., Kavetski, D., Savenije, H. H. G., Clark, M. P., Schoups, G., Pfister, L., and Freer, J.: Catchment properties, function, and conceptual model representation: is there a correspondence?, Hydrol. Process., 28, 2451–2467, doi: 10.1002/hyp.9726, 2014.

Freer, J., Beven, K., and Peters, N.: Multivariate Seasonal Period Model Rejection Within the Generalised Likelihood Uncertainty Estimation Procedure, in Calibration of Watershed Models (eds Q. Duan, H. V. Gupta, S. Sorooshian, A. N. Rousseau and R. Turcotte), American Geophysical Union, Washington, D. C.. 69–87, 2003.

Gupta, H. V., Sorooshian, S., and Yapo, P. O.: Toward improved calibration of hydrologic models: Multiple and noncommensurable measures of information, Water Resour. Res., 34, 751–763, 1998.

Gupta, H. V., Bastidas, L. A., Vrugt, J. A., and Sorooshian, S.: Multiple criteria global optimization for watershed model calibration, Water Sci. Appl., 6, 125–132, 2003.

Gupta, H. V., Clark, M. P., Vrugt, J. A., Abramowitz, G., and Ye, M.: Towards a comprehensive assessment of model structural adequacy, Water Resour. Res., 48, W08301, doi:10.1029/2011WR011044, 2012.

Hock, R.: Temperature index melt modelling in mountain areas, J. Hydrol., 282, 104–115, 2003.

Hublart, P., Ruelland, D., Dezetter, A., and Jourde, H.: Modeling current and future trends in water availability for agriculture on a semi-arid and mountainous Chilean catchment, in: Cold and Mountain Region Hydrological Systems Under Climate Change: Towards Improved Projections, IAHS-AISH P., 360, 26–32, 2013.

Hublart, P., Ruelland, D., Dezetter, A., and Jourde, H.: Assessing the capacity to meet irrigation water needs for viticulture under climate variability in the Chilean Andes, in: Hydrology in a Changing World: Environmental and Human Dimensions, Proc. 7th FRIEND Int. Conf., Montpellier, France, 24–28 February 2014, IAHS-AISH P., 363, 209–214, 2014.

Hublart, P., Ruelland, D., García de Cortázar Atauri, I., and Ibacache, A.: Assessing the reliability of conceptual hydrological modeling in a cultivated, drought-prone catchment of the Chilean Andes, in: Hydrologic Non-Stationarity and Extrapolating Models to Predict the Future, IAHS-AISH P. (in press), 2015.

Huisman, J. A., Rings, J., Vrugt, J. A., Sorg, J., Vereecken, H.: Hydraulic properties of a model dike from coupled Bayesian and multi-criteria hydrogeophysical inversion, J. Hydrol., 380, 62–73, 2010.

IPCC: Full Report: the Physical Science Basis, in: Contribution of Working Group I to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change, Climate Change 2013, edited by: Stocker, T. F., Qin, D., Plattner, G.-K., Tignor, M., Allen, S. K., Boschung, J., Nauels, A., Xia, Y., Bex, V., and Midgley, P. M., Cambridge University Press, Cambridge, UK and New York, NY, USA, 1261–1264, 2013.

Jakeman, A. J. and Hornberger, G. M.: How much complexity is warranted in a rainfall-runoff model?, Water Resour. Res., 29, 2637–2649, 1993.

Jakeman, A. J., and Letcher, R. A.: Integrated assessment and modelling: features, principles and examples for catchment management, Environ. Modell. Softw., 18, 491–501, 2003.

Jothityangkoon, C., Sivapalan, M., and Farmer, D. L.: Process controls of water balance variability in a large semi-arid catchment: downward approach to hydrological model development, J. Hydrol., 254, 174–198, 2001.

Jourde, H., Rochette, R., Blanc, M., Brisset, N., Ruelland, D., Freixas, G., and Oyarzun, R. : Relative contribution of groundwater and surface water fluxes in response to climate variability of a mountainous catchment in the Chilean Andes, in: Cold Regions Hydrology in a Changing Climate, IAHS-AISH P., 346, 180–188, 2011.

Kalthoff, N., Fiebig-Wittmaack, M., Meißner, C., Kohler, M., Uriarte, M., Bischoff-Gauß, I., and Gonzales, E.: The energy balance, evapo-transpiration and nocturnal dew deposition of an arid valley in the Andes, J. Arid Environ., 65, 420–443, 2006.

Kavetski, D., and Kuczera, G.: Model smoothing strategies to remove microscale discontinuities and spurious secondary optima in objective functions in hydrological calibration, Water Resour. Res., 43, W03411, , doi:10.1029/2006WR005195, 2007.

Kavetski, D., and Fenicia, F.: Elements of a flexible approach for conceptual hydrological modeling: 2. Application and experimental insights, Water Resour. Res., 47, W11511, doi:10.1029/2011WR010748, 2011.

Khu, S. T., and Madsen, H.: Multiobjective calibration with Pareto preference ordering: An application to rainfall-runoff model calibration, Water Resour. Res., 41, W03004, 10.1029/2004WR003041, 2005.

Kiptala, J. K., Mul, M. L., Mohamed, Y. A. and van der Zaag, P.: Modelling stream flow and quantifying blue water using a modified STREAM model for a heterogeneous, highly utilized and data-scarce river basin in Africa, Hydrol. Earth Syst. Sci., 18, 2287–2303, 2014.

Kirchner, J. W.: Getting the right answers for the right reasons: Linking measurements, analyses, and models to advance the science of hydrology, Water Resour. Res., 42, WR004362, doi:10.1029/2005WR004362, 2006.

Kokkonen, T. S., and Jakeman, A. J.: A comparison of metric and conceptual approaches in rainfall-runoff modeling and its implications, Water Resour. Res., 37, 2345–2352, 2001.

Krueger, T., Freer, J., Quinton, J. N., Macleod, C. J. A., Bilotta, G. S., Brazier, R. E., Butler, P., and Haygarth, P. M.: Ensemble evaluation of hydrological model hypotheses, Water Resour. Res., 46, W07516. doi:10.1029/2009WR007845, 2010.

Lee, G., Tachikawa, Y., and Takara, K.: Comparison of model structural uncertainty using a multi-objective optimization method, Hydrol. Process., 25, 2642–2653, 2011.

Madsen, H.: Automatic calibration of a conceptual rainfall–runoff model using multiple objectives, J. Hydrol., 235, 276–288, 2000.

MacDonell, S., Kinnard, C., Mölg, T., Nicholson, L., and Abermann, J.: Meteorological drivers of ablation processes on a cold glacier in the semiarid Andes of Chile, The Cryosphere, 7, 1833–1870, doi:10.5194/tc-7-1513-2013, 2013.

McDonnell, J. J., Sivapalan, M., Vaché, K., Dunn, S., Grant, G., Haggerty, R., Hinz, C., Hooper, R., Kirchner, J., Roderick, M. L., Selker, J., and Weiler, M.: Moving beyond heterogeneity and process complexity: A new vision for watershed hydrology, Water Resour. Res., 43, W07301, doi:10.1029/2006WR005467, 2007.

McMillan, H.: Effect of spatial variability and seasonality in soil moisture on drainage thresholds and fluxes in a conceptual hydrological model, Hydrol. Process., 26, 2838–2844, doi: 10.1002/hyp.9396, 2012a.

McMillan, H., Tetzlaff, D., Clark, M., and Soulsby, C.: Do time-variable tracers aid the evaluation of hydrological model structure? A multimodel approach, Water Resour. Res., 48, W05501, doi:10.1029/2011WR011688, 2012b.

Michaud, J., and Sorooshian, S.: Comparison of simple versus complex distributed runoff models on a semi-arid watershed, Water Resour. Res., 30, 593–605, 1994.

Milano, M., Ruelland, D., Dezetter, A., Fabre, J., Ardoin-Bardin, S., and Servat, E.: Modeling the current and future capacity of water resources to meet water demands in the Ebro basin, J. Hydrol., 500, 114–126, 2013.

Minville, M., Brissette, F., and Leconte, R.: Uncertainty of the impact of climate change on the hydrology of a nordic watershed, J. Hydrol., 358, 70–83, 2008.

Montecinos, A. and Aceituno, P.: Seasonality of the ENSO-Related Rainfall Variability in Central Chile and Associated Circulation Anomalies, J. Climate, 16, 281–296, 2003.Moore, R. J.: The PDM rainfall-runoff model, Hydrol. Earth Syst. Sci., 11, 483–499, 2007.

Oudin, L., Hervieu, F., Michel, C., Perrin, C., Andréassian, V., Anctil, F., and Loumagne, C.: Which potential evapotranspiration input for a lumped rainfall–runoff model?: Part 2—Towards a simple and efficient potential evapotranspiration model for rainfall–runoff modelling, J. Hydrol., 303, 290–306, 2005.

Parajka, J., and Blöschl, G.: The value of MODIS snow cover data in validating and calibrating conceptual hydrologic models, J. Hydrol., 358, 240–258, 2008.

Pellicciotti, F., Helbing, J., Rivera, A., Favier, V., Corripio, J., Araos, J., Sicart, J.-E. and Carenzo, M.: A study of the energy balance and melt regime on Juncal Norte Glacier, semi-arid Andes of central Chile, using melt models of different complexity, Hydrol. Process., 22, 3980–3997. doi: 10.1002/hyp.7085, 2008.

Perrin, C., Michel, C., and Andréassian, V.: Improvement of a parsimonious model for streamflow simulation, J. Hydrol., 279, 275–289, 2003.

Pourrier, J., Jourde, H., Kinnard, C., Gascoin, S., and Monnier, S.: Glacier meltwater flow paths and storage in a geomorphologically complex glacial foreland: The case of the Tapado glacier, dry Andes of Chile (30° S), J. Hydrol., 519, 1068–1083, doi:10.1016/j.jhydrol.2014.08.023, 2014.

Quintana, J. M. and Aceituno, P.: Changes in the rainfall regime along the extratropical west coast of South America (Chile): 30–43°S, Atmósfera, 25, 1–22, 2012.

Refsgaard, J. C., and Knudsen, J.: Operational validation and intercomparison of different types of hydrological models, Water Resour. Res., 32, 2189–2202, 1996.

Ruelland, D., Brisset, N., Jourde, H., and Oyarzun, R.: Modelling the impact of climatic variability on the groundwater and surface flows from a mountainous catchment in the Chilean Andes, in: Cold Regions Hydrology in a Changing Climate, IAHS-AISH P., 346, 171–179, 2011.

Ruelland, D., Ardoin-Bardin, S., Collet, L., and Roucou, P.: Simulating future trends in hydrological regime of a large Sudano-Sahelian catchment under climate change, J. Hydrol., 424–425, 207–216, 2012.

Ruelland, D., Dezetter, A., and Hublart, P. : Sensitivity analysis of hydrological modelling to climate forcing in a semi-arid mountainous catchment, in: Hydrology in a Changing World: Environmental and Human Dimensions, Proc. 7th FRIEND Int. Conf., Montpellier, France, 24–28 February 2014, IAHS-AISH P., 363, 145–150, 2014.

Savenije, H. H. G.: HESS Opinions "The art of hydrology", Hydrol. Earth Syst. Sci., 13, 157–161, doi:10.5194/hess-13-157-2009, 2009.

Schaefli, B., Harman, C. J., Sivapalan, M., and Schymanski, S. J.: HESS Opinions: Hydrologic predictions in a changing environment: behavioral modeling, Hydrol. Earth Syst. Sci., 15, 635–646, doi:10.5194/hess-15-635-2011, 2011.

Schreider, S., Whetton, P. H., Jakeman, A. J., and Pittock, A. B.: Runoff modelling for snow-affected catchments in the australian alpine region, eastern Victoria, J. Hydrol., 200, 1–23, 1997.

Seibert, J.: Multi-criteria calibration of a conceptual runoff model using a genetic algorithm, Hydrol. Earth Syst. Sci., 4, 215–224, doi:10.5194/hess-4-215-2000, 2000.

Seibert, J. and McDonnell, J. J.: On the dialog between experimentalist and modeler in catchment hydrology: Use of soft data for multicriteria model calibration, Water Resour. Res., 38, W01241, doi: 10.1029/2001WR000978, 2002.

Seibert, J. and Vis, M. J. P.: Teaching hydrological modeling with a user-friendly catchment-runoff-model software package, Hydrol. Earth Syst. Sci., 16, 3315–3325, 2012.

Shafii, M. and De Smedt, F.: Multi-objective calibration of a distributed hydrological model (WetSpa) using a genetic algorithm, Hydrol. Earth Syst. Sci., 13, 2137–2149, 2009.

Sivapalan, M., Blöschl, G., Zhang, L., and Vertessy, R.: Downward approach to hydrological prediction, Hydrol. Process., 17, 2101–2111, 2003.

Sivapalan, M.: Pattern, process and function: elements of a unified theory of hydrology at the catchment scale, Encyclopedia of Hydrological Sciences, doi:10.1002/0470848944.hsa012, online first, 2006.

Smith, T. J., and Marshall, L. A.: Exploring uncertainty and model predictive performance concepts via a modular snowmelt-runoff modeling framework, Environ. Modell. Softw., 25, 691–701, 2010.

Son, K., and Sivapalan, M.: Improving model structure and reducing parameter uncertainty in conceptual water balance models through the use of auxiliary data, Water Resour. Res., 43, W01415, doi:10.1029/2006WR005032, 2007.

Souvignet, M: Climate Change Impacts on Water Availability in the Semiarid Elqui Valley, Chile, Ph.D. thesis, Cologne University of Applied Sciences, Institute for Technology in the Tropics, 110 pp., 2007.

Souvignet, M., Hartmut, G., Lars, R., Kretschmer, N., and Oyarzún, R.: Statistical downscaling of precipitation and temperature in north-central Chile: an assessment of possible climate change impacts in an arid Andean watershed, Hydrol. Sci. J., 55, 41–57, 2010.

Squeo, F. A., Veit, H., Arancio, G., Gutiérrez, J. R., Arroyo, M. T. K., and Olivares, N.: Spatial heterogeneity of high mountain vegetation in the Andean desert zone of Chile (30ºS), Mt. Res. Dev., 13, 203–209, 1993.

Staudinger, M., Stahl, K., Seibert, J., Clark, M. P., and Tallaksen, L. M.: Comparison of hydrological model structures based on recession and low flow simulations, Hydrol. Earth Syst. Sci., 15, 3447–3459, doi:10.5194/hess-15-3447-2011, 2011.

Strauch, G., Oyarzun, J., Fiebig-Wittmaack, M., González, E., and Weise, S. M.: Contributions of the different water sources to the Elqui river runoff (northern Chile) evaluated by H/O isotopes, Isot. Environ. Health S., 42, 303–322, 2006.

Verbist, K., Robertson, A. W., Cornelis, W. M., and Gabriels, D.: Seasonal predictability of daily rainfall characteristics in central northern Chile for dry-land management, J. Appl. Meteorol. Clim., 49, 1938–1955, 2010.

Vicuña, S., Garreaud, R., and McPhee, J.: Climate change impacts on the hydrology of a snowmelt driven basin in semiarid Chile, Climatic Change, 105, 469–488, 2011.

Wagener, T., Lees, M. J., and Wheater, H. S.: A toolkit for the development and applications of parsimonious hydrological models, in: Mathematical Models of Large Watershed Hydrology, vol. 1, edited by: Singh, V. P. and Frevert, D., Water Resources Publishers, Highland Ranch, CO, 87–136, 2002.

Wainwright, J., and Mulligan, M. (Eds): Environmental modelling – Finding simplicity in complexity. Chichester, John Wiley & Sons, Ltd., 2004.

Xu, C.-Y., and Singh, V. P.: Review on regional water resources assessment models under stationary and changing climate, Water Resour. Manage., 18, 591–612, 2004.

Young, G., Zavala, H., Wandel, J., Smit, B., Salas, S., Jimenez, E., Fiebig, M., Espinoza, R., Diaz, H., and Cepeda, J.: Vulnerability and adaptation in a dryland community of the Elqui Valley, Chile, Climatic Change, 98, 245–276, 2010.

TABLES & CAPTIONS

Table 1. Constitutive equations of fluxes between the various components of the modeling options described in Fig. 2. Parameter (in italic) significations and units are detailed in Table 2. P: catchment-averaged daily precipitation; Rain: rain fraction of precipitation P; Snow: snow fraction of precipitation P; T: catchment-averaged daily temperature; PE: catchment-averaged daily potential evapotranspiration; AE: catchment-averaged daily actual evapotranspiration; $\mathrm{S_j}, \mathrm{j} \in [1{,}5]$: state variables of the conceptual stores; $\mathrm{Q_j}, \mathrm{j} \in [1{,}5]$: water fluxes between the model components).

Options	Constitutive equations	Options	Constitutive equations
A1	$\mathrm{Snow} = \mathrm{P}/(1 + \exp[(\mathrm{T} - \mathrm{T_S})/\mathrm{m_S}])$ $\mathrm{Rain} = \mathrm{P} - \mathrm{Snow}$	C3	$\mathrm{Q_1} = (\mathrm{Melt} + \mathrm{Rain})[1 - (1 - \mathrm{S_1}/S_m)^b]$ $\mathrm{Q_2} = K_1\mathrm{S_1}$
B1a, B1b, B1c	$\mathrm{Melt} = \mathrm{MF}(\overline{\mathrm{T}} - \log[1 + \exp(-\overline{\mathrm{T}})])$ with $\overline{\mathrm{T}} = (\mathrm{T} - T_M)/\mathrm{m_M}$ and $\mathrm{m_M} = 0.1°\mathrm{C}$	D1	$\mathrm{Q_3} = \mathrm{Q_2}$ and $\mathrm{Q_4} = \mathrm{Q_1}$ or $\mathrm{Q_3} = \mathrm{Q_1}$
B1a	$\mathrm{MF} = f_M\mathrm{m_M}$	D2	$\mathrm{Q_3} = \mathrm{Q_1} + \mathrm{Q_2}$
B1b	$\mathrm{MF} = r_1 + r_2\mathrm{T_{30}}$ with $\mathrm{T_{30}}$ the mean temperature of the last 30 days	D3	$\mathrm{Q_3} = (1 - \alpha)\mathrm{Q_1}$ $\mathrm{Q_4} = \alpha\mathrm{Q_1}$
B1c	$\mathrm{MF} = f_1 + f_2\sin(0.551\pi + 2\pi\mathrm{d}/366)$	E1	$\mathrm{Q_{j,lag}} = \mathrm{Q_2}$ with $j \in \{3,4\}$
C1	$\mathrm{AE} = \min(\mathrm{Melt} + \mathrm{Rain},\ K_C\mathrm{PE})$	E2	$\mathrm{Q_{j,lag}}(\mathrm{t}) = \sum_{\mathrm{i}=1}^{N_b} \omega(\mathrm{i})\mathrm{Q_j}(\mathrm{t} - \mathrm{i} + 1)$ with $\omega(\mathrm{i}) = \int_{\mathrm{i}-1}^{\mathrm{i}} 2\mathrm{u}\mathrm{du}/N_b^2$
C2, C3	$\mathrm{AE} = \mathrm{PE}\min(1,\ \mathrm{S_1}/S_m)$	F1a, F2a, F3a	$\mathrm{Q_5} = K_2\mathrm{S_2}^{1+\delta}$ $\mathrm{Q_6} = K_3\mathrm{S_3}$
C1	$\mathrm{Q_1} = \mathrm{Melt} + \mathrm{Rain}$	F1b, F2b, F3b	$\mathrm{Q_5} = K_4\mathrm{S_2} + K_2(\overline{\mathrm{S_2}} - \log[1 + \exp(-\overline{\mathrm{S_2}})])$ $\mathrm{Q_6} = K_3\mathrm{S_3}$ with $\overline{\mathrm{S_2}} = (\mathrm{S_2} - S_C)/\mathrm{m_C}$ and $\mathrm{m_C} = 0.1\ \mathrm{mm^{-1}}$
C2	$\mathrm{Q_1} = (\mathrm{Melt} + \mathrm{Rain})(\mathrm{S_1}/S_m)^\beta$	F3a, F3b	$\mathrm{Q_6} = D\mathrm{S_2}$

Table 2. Parameters used in the various modeling options with their signification and initial sampling. (*) The possible values for K_C were limited to a maximum of 0.5 to reflect the extreme aridity of the catchment.

Parameter	Options	Signification	Units	Initial range
T_S	A1	Rain / snow partitioning temperature threshold	°C	-10 – 10
m_S	A1	Rain / snow partitioning smoothing parameter	–	0.01 – 3
T_M	B1a, B1b, B1c	Snowmelt temperature threshold	°C	-10 – 10
f_M	B1a	Constant melt factor	$°C.mm^{-1}$	0 – 10
r_1	B1b	Coefficient for computation of the variable melt factor	$°C.mm^{-1}$	1 – 5
r_2	B1b	Coefficient for computation of the variable melt factor	$°C.mm^{-1}$	1 – 5
f_1	B1c	Coefficient for computation of the variable melt factor	$°C.mm^{-1}$	1 – 5
f_2	B1c	Coefficient for computation of the variable melt factor	$°C.mm^{-1}$	1 – 5
K_C	C1	Evapotranspiration coefficient	–	0.05 – 0.5 (*)
S_m	C2, C3	Maximum storage capacity of the moisture-accounting store	mm	10 – 100
β	C2	Shape parameter	–	0.1 – 3
b	C3	Shape parameter of Pareto distribution	–	0.1 – 3
K_1	C3	Infiltration coefficient	d^{-1}	0.001 – 0.7
α	D3	Splitting parameter	–	0.1 – 0.9
N_b	E2	Number of time steps in the lag routine	–	1 – 6
K_2	F1a to F3b	Storage coefficient	d^{-1}	0.01 – 0.99
K_3	F1a to F3b	Storage coefficient	d^{-1}	0.001 – 0.01 (F1a, F1b, F3a, F3b) 0.001 – 0.1 (F2a, F2b)
δ	F1a, F2a, F3a	Power law parameter of the non-linear store in the runoff transformation module	–	0 – 1
S_c	F1b, F2b, F3b	Threshold parameter of the non-linear store in the runoff transformation module	mm	10 – 300
D	F3a, F3b	Recharge coefficient	d^{-1}	0.001 – 0.5
K_4	F1b, F2b, F3b	Storage coefficient	d^{-1}	0.001 – 0.01

Table 3. Coordinates of the cluster centroids in the four-dimensional (4D) space of performance measures. The number of models with membership values > 50% ($N_{50\%}$) is given for each cluster.

Calibration period (1997–2011)

Cluster no.	Crit1 (1-NSE)	Crit2 (1-NSE_{log})	Crit3 (VE_M) (%)	Crit4 (SE) (%)	$N_{50\%}$
1	0.15	0.25	10	9	24
2	0.23	0.30	10	10	24
3	0.49	0.58	23	11	10
4	0.60	0.62	25	16	13
5	0.92	0.97	33	20	1

Validation period (1982–1996)

Cluster no.	Crit1 (1-NSE)	Crit2 (1-NSE_{log})	Crit3 (VE_M) (%)	Crit4 (VE_C) (%)	$N_{50\%}$
1	0.24	0.21	14	3	15
2	0.32	0.29	15	4	25
3	0.38	0.31	15	5	8
4	0.51	0.42	25	23	8
5	0.61	0.44	27	27	11
6	0.61	0.51	30	33	5

Table 4. Detailed composition of Clusters 1 in calibration and validation. The tables indicate the numbers and the names of the models as well as their number of parameters NP. For each criterion only the best performance value obtained along the Pareto front is given. N_{par} (%) represents the proportion of observations enclosed within the simulation bounds of each Pareto set of solutions. Asterisks are used to indicate the models which are not in the best-performing group (Cluster 1) either in calibration or in validation.

Calibration period (1997–2011)

Model no.	Model name (options)	NP	NSE	NSE_{log}	VE_M (%)	SE (%)	N_{Par} (%)
2	A1–B1a–C1–D1–E1–F2b	9	0.87	0.76	10.6	11.2	76.0
4	A1–B1a–C1–D1–E1–F3b	10	0.84	0.77	10.4	11.2	53.2
8	A1–B1a–C1–D3–E2–F2b	11	0.83	0.75	11.7	11.1	76.5
20	A1–B1a–C3–D1–E2–F2b	12	0.83	0.76	10.0	11.4	60.0
22	A1–B1a–C3–D2–E1–F2b	11	0.90	0.77	10.4	11.2	64.1
26	A1–B1b–C1–D1–E1–F2b	10	0.87	0.77	10.1	11.5	58.4
30 (*)	A1–B1b–C1–D3–E2–F1b	12	0.84	0.70	9.8	11.4	69.6
32 (*)	A1–B1b–C1–D3–E2–F2b	12	0.83	0.71	11.1	11.4	68.4
44	A1–B1b–C3–D1–E2–F2b	13	0.89	0.77	10.6	11.4	63.4
46	A1–B1b–C3–D2–E1–F2b	12	0.90	0.76	10.7	11.4	45.4
49 (*)	A1–B1c–C1–D1–E1–F2a	9	0.82	0.73	10.9	7.0	67.0
50	A1–B1c–C1–D1–E1–F2b	10	0.86	0.77	10.4	7.0	67.4
52 (*)	A1–B1c–C1–D1–E1–F3b	11	0.85	0.72	8.8	8.1	65.7
53 (*)	A1–B1c–C1–D3–E2–F1a	11	0.79	0.76	10.8	7.0	63.8
54	A1–B1c–C1–D3–E2–F1b	12	0.90	0.78	11.5	7.5	55.7
55 (*)	A1–B1c-C1-D3–E2–F2a	11	0.80	0.73	10.7	7.0	54.5
56	A1–B1c–C1–D3–E2–F2b	12	0.85	0.75	10.8	7.6	76.3
65	A1–B1c–C3–D1–E2–F1a	12	0.83	0.78	8.0	7.7	65.0
66 (*)	A1–B1c–C3–D1–E2–F1b	13	0.81	0.77	9.6	6.8	63.5
67 (*)	A1–B1c–C3–D1–E2–F2a	12	0.81	0.75	10.7	7.0	73.7
68	A1–B1c–C3–D1–E2–F2b	13	0.85	0.74	10.6	6.8	74.5
69 (*)	A1–B1c–C3–D2–E1–F2a	11	0.82	0.73	10.6	7.0	51.8
70	A1–B1c–C3–D2–E1–F2b	12	0.87	0.76	10.7	7.5	76.4
72 (*)	A1–B1c–C3–D2–E1–F3b	13	0.81	0.71	9.8	7.1	69.0

Validation period (1982–1996)

Model no.	Model name	NP	NSE	NSE_{log}	VE_M (%)	VE_C (%)	N_{Par} (%)
2	A1–B1a–C1–D1–E1–F2b	9	0.75	0.78	13.3	2.7	87.1
4	A1–B1a–C1–D1–E1–F3b	10	0.73	0.80	14.1	3.8	50.0
8	A1–B1a–C1–D3–E2–F2b	11	0.75	0.76	14.5	5.8	84.8
20	A1–B1a–C3–D1–E2–F2b	12	0.72	0.77	13.7	3.7	58.4
22	A1–B1a–C3–D2–E1–F2b	11	0.76	0.78	12.3	3.3	75.3
26	A1–B1b–C1–D1–E1–F2b	10	0.74	0.78	12.9	3.5	70.2
42 (*)	A1–B1b–C3–D1–E2–F1b	13	0.73	0.75	15.6	3.3	62.7
44	A1–B1b–C3–D1–E2–F2b	13	0.74	0.79	13.0	4.1	69.3
46	A1–B1b-C3–D2–E1–F2b	12	0.76	0.77	15.2	3.4	48.4
50	A1–B1c–C1–D1–E1–F2b	10	0.78	0.81	13.9	2.5	73.1
54	A1–B1c–C1–D3–E2–F1b	12	0.77	0.78	15.3	3.5	60.8
56	A1–B1c–C1–D3–E2–F2b	12	0.75	0.77	13.2	4.5	81.3
65	A1–B1c–C3–D1–E2–F1a	12	0.74	0.80	13.8	3.6	73.0
68	A1–B1c–C3–D1–E2–F2b	13	0.77	0.74	13.5	3.7	78.7
70	A1–B1c–C3–D2–E1–F2b	12	0.73	0.78	14.2	3.4	79.4

FIGURES & CAPTIONS

Figure 1. The Claro River Basin at Rivadavia (1515 km²) in Chile: topography and mean annual precipitation and temperature over 1982–2011 (based on Ruelland *et al.*, 2014). Several of the stations used in this study were located outside the catchment and therefore not displayed on the following maps.

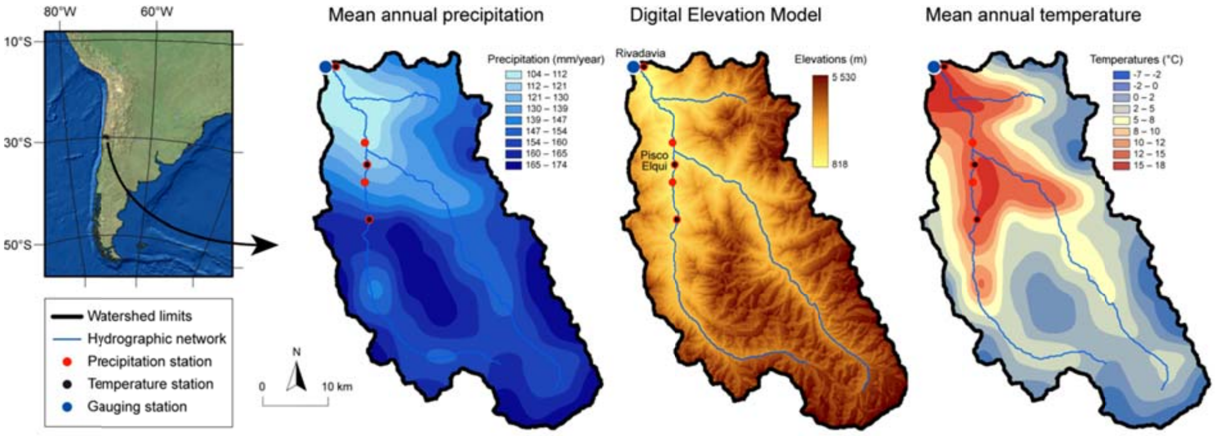

Figure 2. Interannual variability in precipitation and observed streamflow from 1989 to 2008. The hydrological year was defined from May to April so as to capture the snowmelt and peak flow seasons at mid-year. Streamflow values are those measured at the catchment outlet before accounting for water abstractions. Precipitation values are those obtained after interpolation.

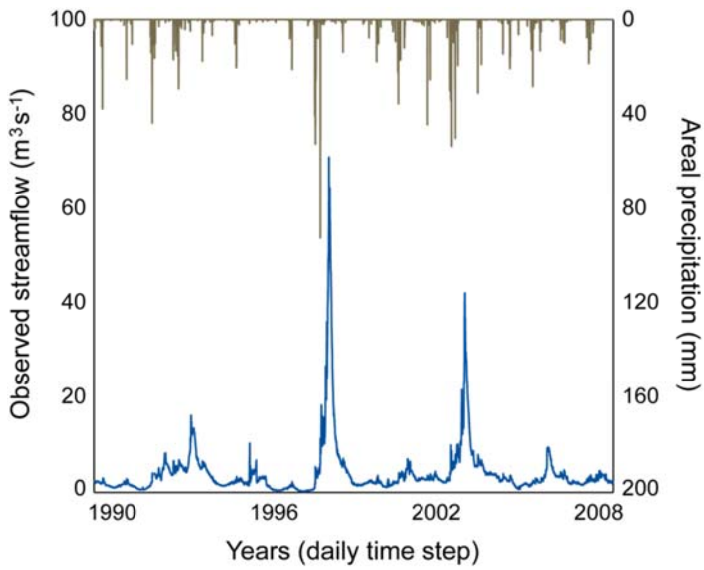

Figure 3. Overall architecture (modules), decision tree and available modeling options of the modular multiple-hypothesis framework (P: catchment-averaged daily precipitation; SWE: snow water equivalent; AE: catchment-averaged daily actual evapotranspiration; $S_j, j \in [1,5]$: state variables of the conceptual stores; $Q_j, j \in [1,5]$: water fluxes between the model components).

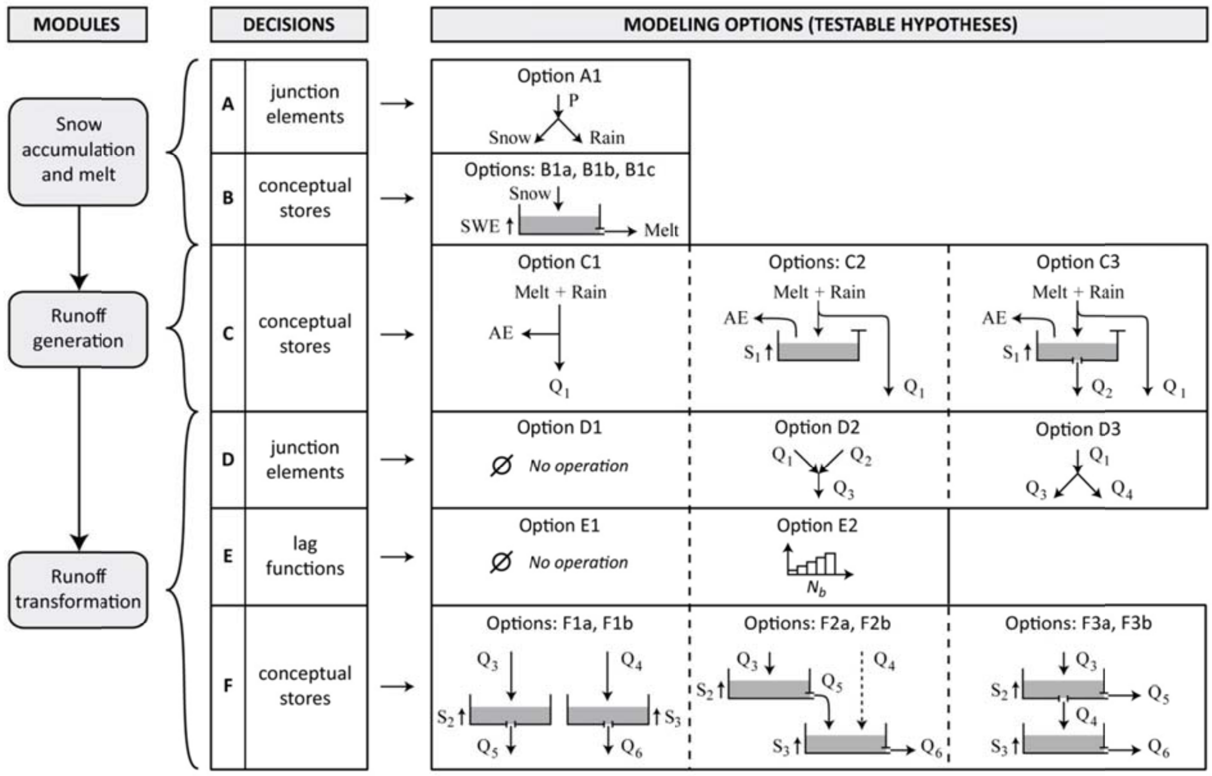

Figure 4. Description of the snow error criterion. The overall snow error (SE) can be described as a sum of two terms, SE1 and SE2, whose values are given by a confusion matrix. In this example, water storage in the snow-accounting store (solid line) starts (SE1) and ends (SE2) sooner than what would be expected from the SCA data (dashed line).

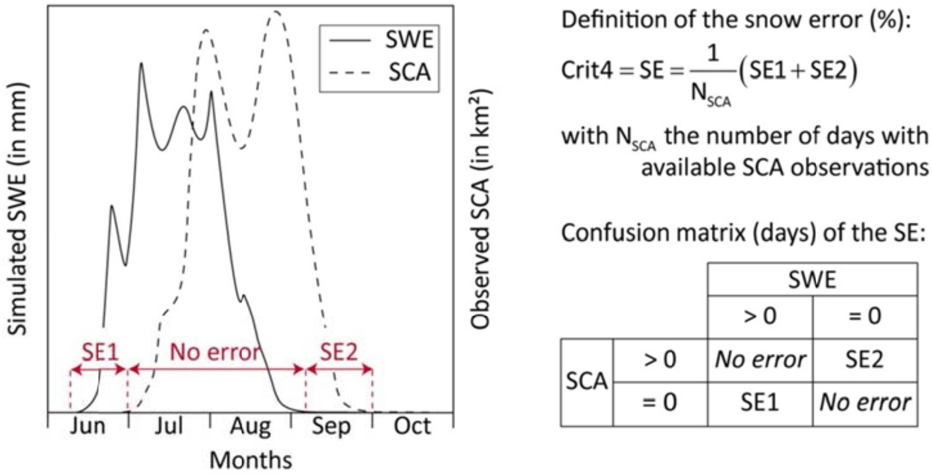

Figure 5. Projections of the clusters onto three possible planes of the objective space in calibration and validation. As explained in Sect 3.3., each point represents a different model hypothesis.

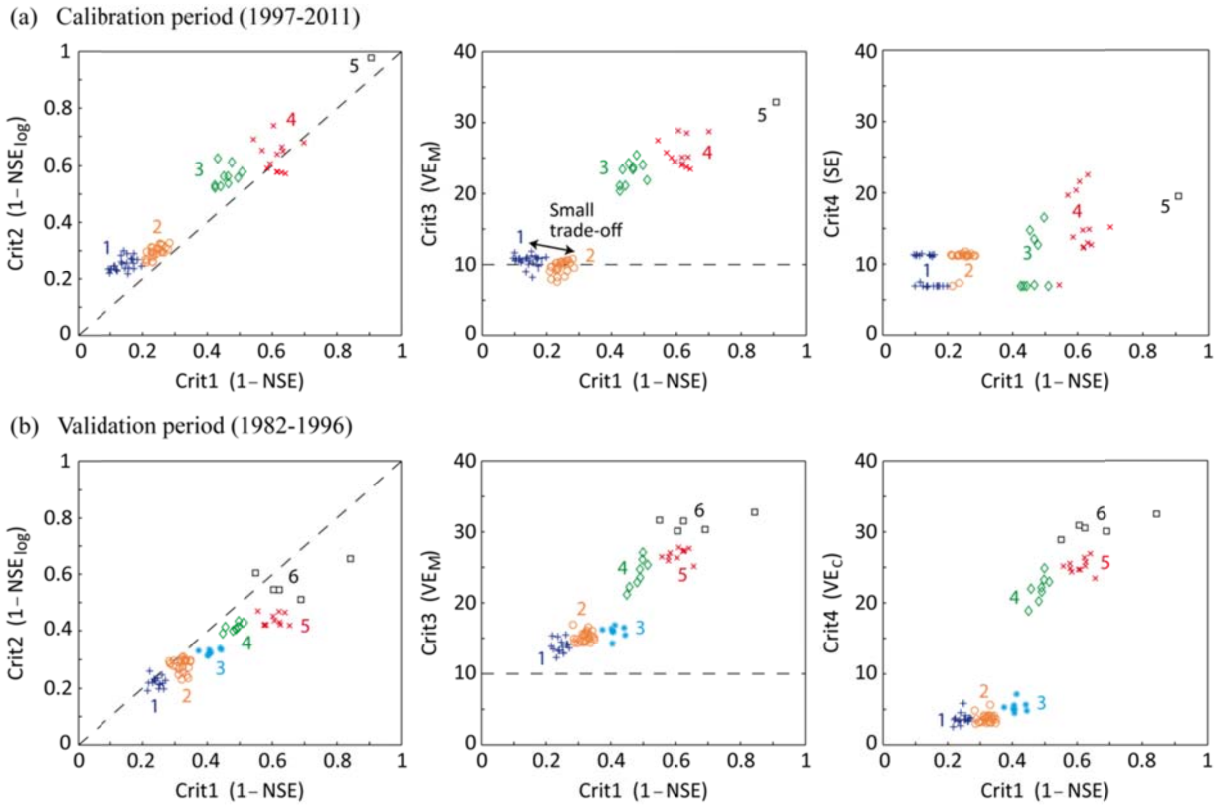

Figure 6. Projections of the Pareto fronts of model hypotheses (a) no. 49 (A1-B1c-C1-D1-E1-F2a) and (b) no. 50 (A1-B1c-C1-D1-E1-F2b) onto three possible two-dimensional subspaces of the objective space.

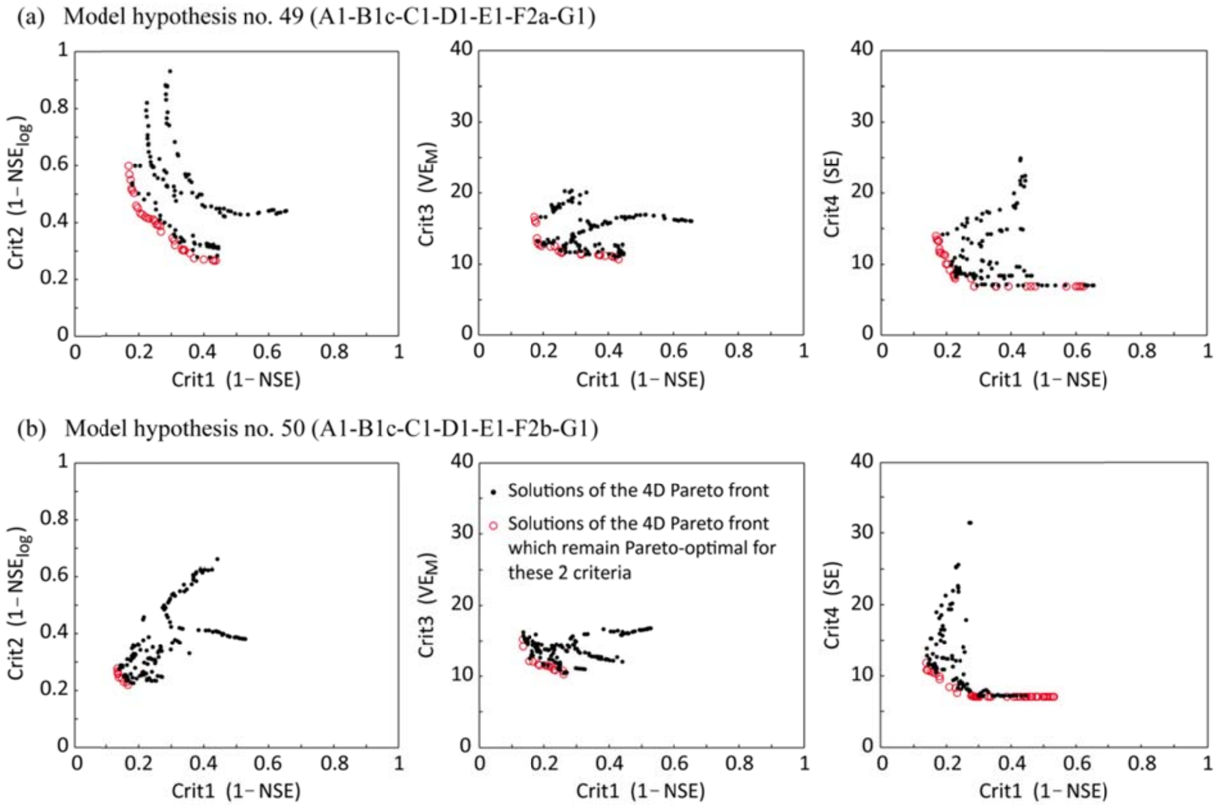

Figure 7. Estimated normalized ranges of the Pareto-optimal sets of eight alternative model structures differing in at least one of their components. The colored lines stand for the best solutions obtained in calibration with respect to the high flow criterion (in black), the low flow criterion (in red), the mean annual volume error (in blue) and the snow error (in green).

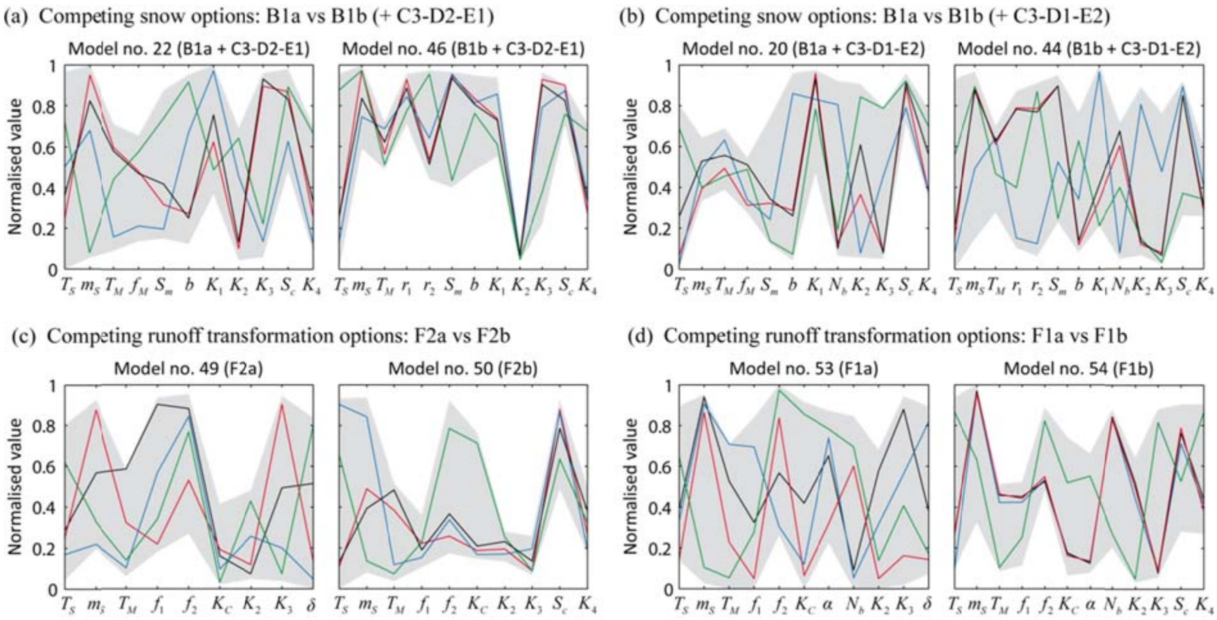

Figure 8. Comparison of MODIS-based SCA data (red dashed lines) with the SWE simulations (shaded areas) of models no. 6, 30 and 54. The shaded area corresponds to the range of SWE simulations obtained from the Pareto sets of these models.

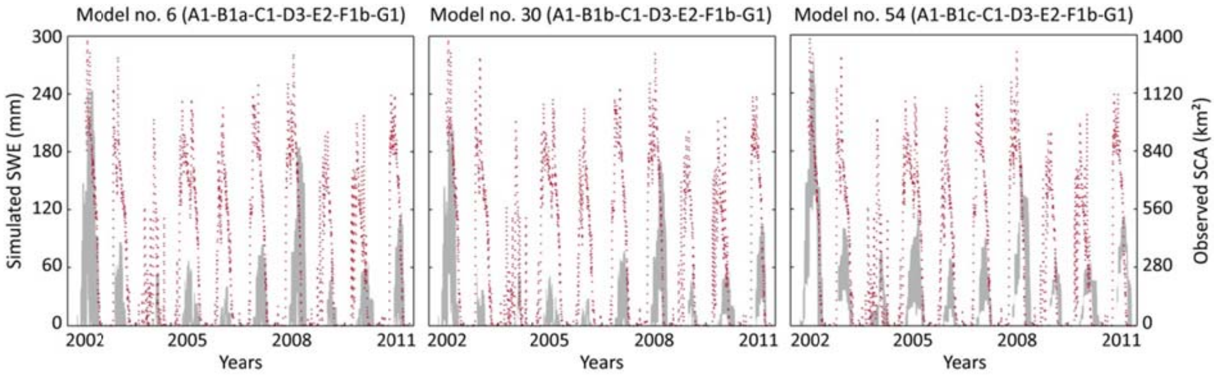

Figure 9. Comparison of observed daily discharge at Rivadavia with the overall uncertainty envelope obtained by combining the Pareto-envelopes of 8 model structures. These structures have been selected among the 14 members of Cluster 1 in both calibration and validation so as to minimize the uncertainty envelope area (Area, in pixels^2) while holding constant the number of outlying observations (Outlying, in %). The red parts indicate potential errors in the model structures or observed data.