Optimization of the Hydrologic Response Units (HRU) using gridded meteorological data and spatially varying parameters

Enter authors here: D. Poblete[1]†, J. Arevalo [2,3], O. Nicolis [4] and Felipe Figueroa[5]

[1] Escuela de Ingeniería Civil, Universidad de Valparaíso, Valparaíso, Chile.

[2] Departamento de Meteorología, Universidad de Valparaíso, Valparaíso, Chile.

[3] Department of Hydrology and Atmospheric Sciences, University of Arizona, Tucson, AZ, USA.

[4] Facultad de Ingeniería, Universidad Andrés Bello, Viña del Mar, Chile.

[5] Plataforma de Investigación en Ecohidrología y Ecohidráulica (EcoHyd), Santiago, Chile.

Corresponding author: David Poblete (david.poblete@uv.cl)

† David Poblete, email address: david.poblete@uv.cl

Key Points:

- The paper presents a methodology to optimize the definition of Hydrologic Response Units (HRUs) for semi-distributed hydrologic modelling.
- The optimization minimizes the internal variability within an HRU and maximize the variance between different HRUs.
- The results show different hydrological responses by each HRU, in terms of total volume, timing, distribution and peak discharge.

Abstract

Although complex hydrological models with detailed physics are every day more common, lumped and semi-distributed models are still used for many applications and offer some advantages as its reduced computational cost. Most of these semi-distributed models use the concept of Hydrological Response Unit or HRU. In its original conception, HRUs are defined as homogeneous structured elements having similar climate, land-use, soil and/or pedotransfer properties, hence a homogeneous hydrological response under equivalent meteorological forcing. This work presents a quantitative methodology to construct HRUs based on Principal Component Analysis and Hierarchical Cluster Analysis of gridded meteorological data and hydrological parameters. The methodology is tested using the Water Evaluation and Planning System (WEAP) model for the Alicahue River Basin, a small catchment in Central Andes, in Central Chile. The results show that with four HRUs it is possible to reduce up to about a 10% the relative within variance of the catchment, an indicator of homogeneity of the HRUs. Evaluation of the simulations show a good agreement with streamflow observations in the outlet of the catchment with a NSE value of 0.79 and also show the presence of small hydrological extreme areas that generally are neglected due to their relative size.

Plain Language Summary

1 Introduction

Since the works of Leavesley et al. (1983) and Flügel (1995), the concept of Hydrologic Response Units (HRU) has risen as one of the most common approaches for semi-distributed hydrological modelling. Flügel (1995) defined an HRU as a homogeneous structured element having similar climate, land-use, soil and/or pedotransfer properties, hence a homogeneous hydrological response under equivalent meteorological forcing. An important assumption is that the variation of the hydrological process dynamics within a single HRU is small compared with the hydrologic dynamics and responses to other units defined in the model. Many authors assume that HRU do not necessarily represent contiguous geographical areas so the topology of the elements is simplified or just neglected (Pilz, Francke, & Bronstert, 2017; Savvidou, Efstratiadis, Koussis, Koukouvinos, & Skarlatos, 2018) and the total discharge of the watershed is calculated as the incremental input of every independent element and propagated to its outlet; assumption that we will also consider in the rest of this study.

Traditionally, land use/land cover, topographic characteristics and soil types have been used as proxies of many of the parameters involved in the governing equations and parameterizations of the lumped, semi-distributed and even distributed models, but always with a certain degree of uncertainties (Höge, Wöhling, & Nowak, 2018; Nijzink et al., 2016; Orth, Dutra, & Pappenberger, 2016). Most of the methodologies to delineate HRUs rest on the expected relationships between physical-ecological characteristics of the catchment and the corresponding hydrological properties reflected on the hydrological model parameters. Hence, HRUs are usually defined by the superposition of land use and soil type and after the classification, quantitative or qualitative relations are used to estimate hydrologic parameters on each HRU. One of the most common approaches has been to include the sub-basins in the process, hence the

intersection of the sub-basins, land use categories and soil type polygons in a GIS represents the minor elements for hydrologic modelling (Dehotin & Braud, 2008; Savvidou et al., 2018).
A different approach is used in Savvidou et al. (2018), as they estimate the CN Curve Number parameter for reference conditions using soil permeability, vegetation classes and drainage capacity maps and then the HRUs are defined based on the separation of areas according to the CN values. According to the authors, this delineated HRUs can be used in any hydrological model as the SCS-CN model, which is widely used and understood.
Even though the importance on defining properly the HRU for a good representation of the hydrological processes and dynamics, methods and tools for identifying an appropriate scale, are often missing. The challenge is to identify a proper method for discretization of the basins, losing the least information possible and maximizing the model reliability and utility that in turn plays a crucial role in the accuracy of the models (Haghnegahdar, Tolson, Craig, & Paya, 2015; Han, Huang, Zhang, Li, & Li, 2014; Haverkamp, Fohrer, & Frede, 2005). If over simplification of the basin characteristics is done, small areas of extreme hydrologic behavior can be neglected by a lack of representation in the aggregation procedures (Haverkamp et al., 2005). On the other hand, if the used data is highly detailed and fragmented, it can lead to an excessive number of HRUs, making the modelling impracticable.
Although meteorological variables are inputs to every model, none of the methodologies use that information directly in the construction process of the HRU. Flügel (1995) suggested more than two decades ago that the use of meteorological information to construct HRU is advisable, but it has not been explored in depth probably due to the lack of good quality spatial meteorological information. Today, this idea is more plausible and can be considered because one of the basic assumptions on HRU is that meteorological forcing is homogeneously spatialized over the domain of the HRU. Therefore, the spatial heterogeneity of the precipitation and other variables can be incorporated in the delineation of HRUs. An indirect approach to include climate information is used by Young et al. (2009), where 15 watersheds of the Sierra Nevada in California are discretized in HRU by the intersection of sub-basins, soils type, vegetation cover and elevation bands in the Water Evaluation And Planning System model (WEAP; Yates, Sieber, Purkey & Huber-Lee (2005)). They calculate fractional areas for each sub-basin using a vegetation cover/soil type combination in 250 meters elevation bands ranging from 500 to 4000 meters above sea level, in order to provide a finer discretization for snow accumulation and melt modelling. This has been a common practice in the use of this model in semi-arid basins in Chile (for instance, Bonelli, Vicuña, Meza, Gironás, & Barton (2014) and Vicuña, Garreaud, & McPhee (2010)).
Given all these issues, some questions arise: How to use the detailed information available on land use, geomorphologic properties and climatic behavior for the separation of a manageable number of independent modelling units? Which criteria must be used to simplify the complexity of hydrologic dynamics of a watershed into the smallest number of homogeneous units as possible without losing valuable information? Does the use of these independent modeling units ensure heterogeneity of hydrologic response between them?
This paper presents a quantitative methodology for the determination of unstructured HRU based on homogeneity of the hydrological parameters used by any specific hydrological model and its meteorological inputs. A Principal Component Analysis (PCA) is performed in order to get an independent set of vectors to be used in a Hierarchical Clustering (HC) algorithm to obtain the desired HRUs. The result minimizes the internal variability of hydrologic properties in each

HRU and simultaneously maximizes the variability between different HRUs, subsequently of the hydrologic responses of each element.
To test the proposed methodology, HRU delineation is performed for the Alicahue river basin, an Andean semi-arid basin located in Central Chile. Hydrologic parameters and climate averaged values used by the semi-distributed WEAP model (Water Evaluation and Planning System) are calculated over a regular grid, that in turn are used to classify each cell in the mentioned HRUs. Climate variables are based on a 1km resolution bias-corrected model output for three periods of 12-month using the WRF model (Skamarock et al., 2008) and the hydrologic parameters are estimated by topographic characteristics derived from 30m ASTER DEM (Tachikawa et al., 2011) and Land Use data from Natural Resources Research Center of Chile (Martínez, Flores, Retamal, Ahumada, & Brito, 2013). Finally, the performance, accuracy and skill of the model using ten different configurations of HRU are analyzed using common modelling indicators.

2 Materials and Methods

The proposed methodology to delineate HRUs consists in the creation of a dataset of raster files comprised of hydrologic parameters and meteorological variables used by the target hydrological model. Then, through principal component and hierarchical cluster analyses, every cell of the raster files is classified into a specific cluster to form the different HRUs.

2.1 Area of the study case

Central Chile has a landscape with a very complex topography. It is surrounded by the high peaks of the Andes Mountains usually above 4,000 m.a.s.l. at the East and the Pacific Ocean at only about 150 km West of the mountains. Most of the river basins in this area have a latitudinal preferential path, downstream of the Andes up to the Ocean. Its climate corresponds to Mediterranean with dry summers, temperatures are usually mild ranging from about 0°C as minimum during winter up to 35°C as maximum during summer, except for the high elevation lands where below freezing temperatures are usual during winter. Mean annual precipitation is about 400 mm for the valleys and coastal areas, which is mostly due to winter frontal storms, hence the spatial variability is mostly modulated by the orographic effects.

The domain of the area of study corresponds to the Alicahue river basin, which is located between geographical coordinates 32.39°S to 32.21°S in latitude and 70.76°E to 70.41°E in longitude, in the province of Petorca, Valparaiso Region, Central Chile (Figure 1). This is a sub-catchment of the La Ligua river basin, that receives water from other minor streams and flows into the sea, with a total length of nearly 200km. The Alicahue river has a length of just 30 km and its drainage area is just 354 km2, but its topography ranges from 780 m.a.s.l up to 3985 m.a.s.l., with almost half of its area located above 2500 m.a.s.l. In winter and during rainfall, the 0°C isotherm in central Chile is typically located at about 2500 m.a.s.l. (Garreaud, 1992), allowing snow accumulation in most part of the Andes mountains. Hence, at the outlet of the basin there is an important flow between mid-spring and beginning of summer in southern hemisphere (from October to January) due to snow melting. Agriculture uses the waters from La Ligua river, but most of the discharge of this river during the dry season comes from upper basins as the Alicahue river, where snow accumulation is possible during the cold and wet winter.

The Alicahue river basin is a small catchment and has limited human intervention, which simplifies the analysis of the results. Also, its behavior is comparable to several other high-altitude catchments in Central Chile, where complex topography is dominated for the traversal valleys downstream of the Andes Mountains and snowmelt is one of the dominant hydrologic drivers.

Figure 1. Area of Study. Relative location in South America and WRF Domains (left), topography of the Area of Study indicating the limits of Alicahue and La Ligua river basins (red and purple polygons, respectively) and the location of the weather stations (streamflow station co-located with Alicahue Hacienda weather station).

The only available stream gauge station is located in the outlet of the Alicahue river basin (32.20ºS, 70.45ºW), from station BN 05200001-7 "*Rio Alicahue en Colliguay*", from the General Dictatorate of Water (DGA in Spanish), with a recording period starting the year 1963.

Although only the station Alicahue Hacienda is located inside the basin, the frontal nature of the precipitation makes reasonably to correlate near observations. The values recorded at these stations are 1.16 m3s^{-1} for mean annual streamflow, 267mm for total annual precipitation and 15.1ºC for mean annual temperature.

2.2 Hydrologic Parameters and meteorological datasets

The methodology uses raster maps of the hydrologic parameters and mean annual values of the meteorological variables used by the chosen hydrologic model. These maps need to be constructed or generated previously by any methodology.

In the case of this study, the WEAP model (Yates et al., 2005) is used to test the methodology. WEAP is a water allocation model, that has been used for water resources management in several studies over several catchments around the world (for instance Purkey et al., 2007; Young et al., 2009) and particularly in Chile (Bonelli et al., 2014; Vicuña et al., 2010). It has incorporated a hydrology module that represents the mass balance in elements called *catchments*, in which simplified hydrological fluxes and storages are modelled using a one dimensional and 2-layers storage system. Although a WEAP catchment can be used as a single HRU, the *catchment* element can be internally divided in more separate units, each of them as a single HRU. The methodology most widely used, divides the catchment in elements by land use cover within a given elevation band and sub-basin; with all the HRU having the same meteorological condition on each elevation band.

The upper layer of the *catchment* element has four hydrologic parameters: *Sw* (Soil Water Capacity in mm) represents the soil layer depth; *RRF* (Runoff Resistance Factor) is equivalent to the run-off coefficient in the rational equation; *Ks* (Root Zone Conductivity in mm/month) corresponds to the saturated hydraulic conductivity in the soil layer; and *kc* to the crop coefficient of the vegetation. The parameter *f* (Preferred Flow Direction) controls the water flowing from the upper layer to the lower layer as Interflow or Deep Percolation (*f*=1 for total horizontal flow and *f*=0 for total vertical flow). *Dw* and *kd* represent the depth and the saturated hydraulic conductivity of the deeper layer of the *catchment* element, respectively. Finally, the simple snow model uses two temperature thresholds, for melting and freezing (T_l and T_s, respectively), totalizing nine parameters (for detailed information on the water balance equations see Yates et al., 2005).

Figure 2 shows the maps of the WEAP parameters *f*, *RRF* and the log values of *Sw* and *Ks*, respectively for the Alicahue river basin. *RRF* (top-left) is related to milder slopes and vegetated terrain; *f* (top-right) depends on the terrain slope and soil properties. *Sw* (bottom-left) and *Ks* (bottom-right) are shown in a log scale for better visualization. *Sw* present the deepest soils in the flat lands near the main water course, in contrast to the sides of the hillslope. The latter is also related to the vegetation land cover and slope and dominated by very permeable areas in high altitude wetlands.

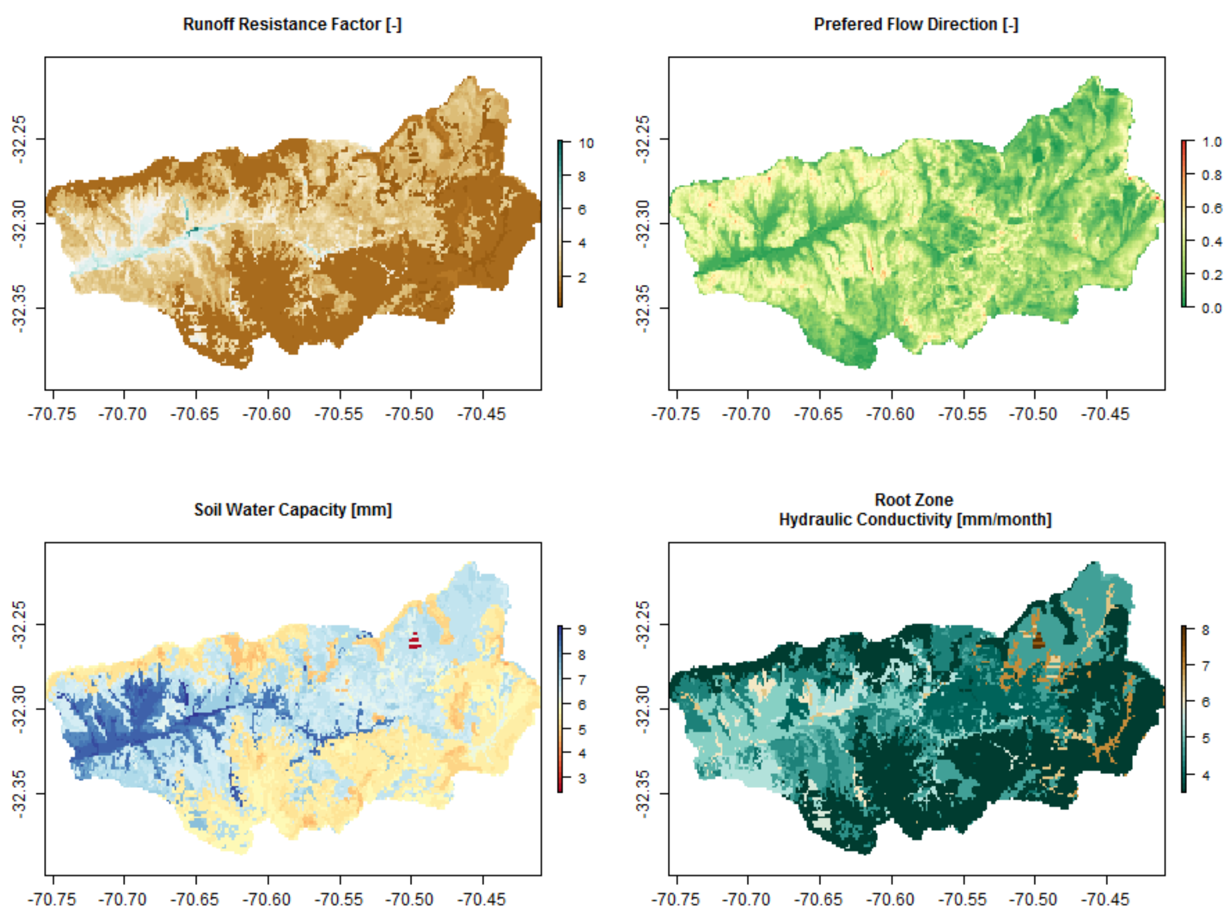

Figure 2. Parameters map covering the Alicahue River basin for: (top left) f, (top right) RRF, (bottom left) Sw and (bottom rigth) Ks. Sw and Ks are plotted in log scale.

As the spatial representativeness of meteorological stations is small in complex terrain and observations are usually scarce, the WRF model version 3.4.1 was used to simulate three periods of 12 consecutive months each: (1) 1 July 2003 to 30 June 2004, (2) 1 June 2009 to 31 May 2010, and (3) 1 Jun 2010 to 31 May 2011. Those simulations were performed in four nested domains with 27, 9, 3 and 1 km of horizontal resolution and 50 terrain-following vertical levels, the innermost domain was used for the analysis and cover entirely the study area, as could be seen in Figure 1. As the WRF simulation period covered only 36 non continuous months, it was not suitable to drive the long-term hydrological modeling. Hence, a simpler relation between the available observed precipitation time series near the basin and thus in each HRU was used. Long-term temperature time series were extrapolated to the HRUs using a simple linear model between the mean annual temperature modeled in WRF and "Alicahue Hacienda" station records using variables as elevation, aspect, mean longitude and mean latitude for the HRUs in each of the simulations.

Figure 3 shows the mean annual precipitation and the mean temperature for the 36 months of WRF simulations. Other climatological variables as relative humidity, net radiation, albedo, evapotranspiration and wind speed are not shown.

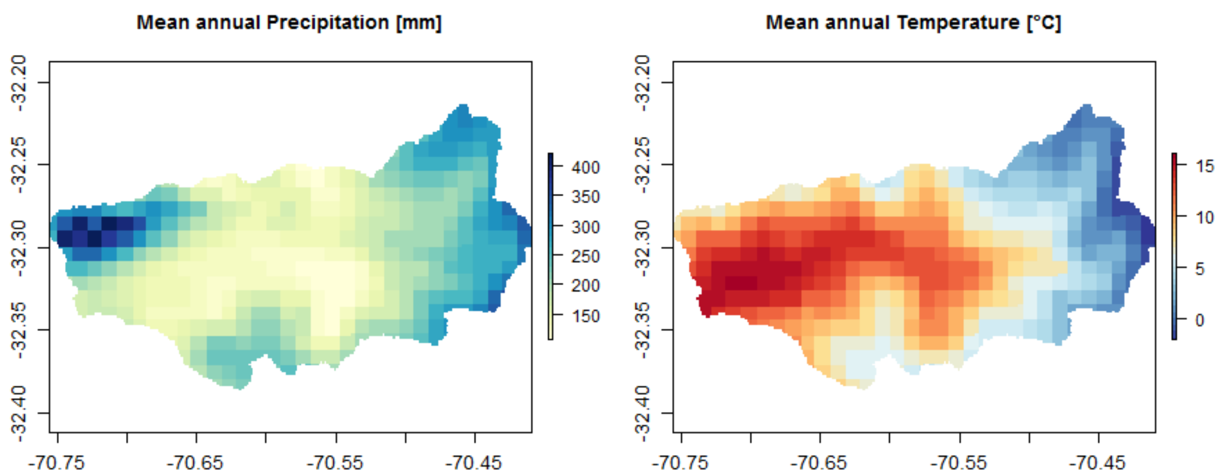

Figure 3. Maps covering the Alicahue River basin for: (left) mean annual precipitation, (right) mean temperature.

As both types of datasets have different cell sizes and extension, to join both datasets the meteorological raster maps are resampled to a common grid system into the parameters base grid by the nearest neighbor method using the Vincenty (ellipsoid) great circle distance from the *distm* function of the geosphere package in R (Hijmans, 2017).

2.3 Clustering processes and HRU delineation

In this section, the core of the HRU delineation process is detailed. The gridded model-specific parameters and the climatological information are used in the Principal Component Analysis to later use its firsts components in the Hierarchical Clustering.

The Principal Component Analysis (PCA) technique consists on describing a multidimensional data set using a smaller number of uncorrelated variables (principal components) that incorporate as much information as possible. If the data set X is composed by i individuals and p variables, the PCA finds the first k principal components (with k<p) with the maximum variance. The Principal Component Analysis is performed using the function PCA from the FactoMineR package (Lê, Josse, & Husson, 2008) for multivariate data analysis in R (R Core Team, 2019) by assigning greater weights to the most important variables, in order to capture more variance of these variables. In this study, precipitation and temperature variables were weighted by a factor of two given its importance in the water balance equation, while the rest of the variables had weights equal to one. This allows the use of the expertise of the modeler in assigning more importance to specific variables.

Working with principal components instead on the original data, allows to obtain more stable results in the clustering process. Since the first dimensions (or components) extract the most information from data and the last ones represent the noise (Husson, Julie, & Jérôme, 2010), the first components accounting at least 90% of variance are used in the Hierarchical Cluster Analysis function HCPC from the same FactoMineR package. The objective is to capture most of the variability of the most important variables and simultaneously not to capture the variability of the least important variables or represent a minor proportion of the main variables.

The Hierarchical Clustering used in this work has been implemented using the Ward's criterion Husson et al. (2010). The Ward's method is based on an agglomerative approach or "bottom-up", where the clustering starts considering each observation as a cluster, and pairs of clusters are merged as one moves up in the hierarchy. The initial cluster distances in Ward's method may be defined by the squared Euclidean distance between the individuals' values and their averages.

By considering a multivariate database composed by *i* spatial individuals (cells) and *K* variables (both hydrologic parameters and meteorological variables), the total variance of *Q* clusters (with *Q*<*i*) is evaluated according to its decomposition in the between and within variances given by:

$$\sum_{k=1}^{K}\sum_{q=1}^{Q}\sum_{i=1}^{N_q}(x_{iqk}-\overline{x}_k)^2 = \sum_{k=1}^{K}\sum_{q=1}^{Q}N_q(\overline{x}_{qk}-\overline{x}_k)^2 + \sum_{k=1}^{K}\sum_{q=1}^{Q}\sum_{i=1}^{N_q}(x_{iqk}-\overline{x}_{qk})^2, \qquad (1)$$

where x_{iqk} is the normalized value of the variable k for the individual i of the cluster q, $\overline{x}_{qk}$ is the mean of the variable k for cluster q, $\overline{x}_k$ is the overall mean of variable k (equal to zero if normalized) and N_q is the number of spatial points in cluster q. The first member at the right side of the equation represents the between inertia (or between variance) and the second member, the within inertia (within variance).

The importance on this equation is that the total variance of the system remains constant and as the within variance decreases (the clusters become more homogeneous), the between variance increases (the clusters become more and more different between each other).

At each step of the aggregating procedures algorithm the increase of within variance is minimized (or the increase of the between variance is maximized). This analysis detects groups of individuals with similar characteristics and hydrologic behavior based on parameters and meteorological similitude between cells. Each group of cells belonging to a cluster represents a single HRU to be used in the hydrological model. It is not necessary for cells to be contiguous to belong to the same cluster as proximity in the space of attributes does not ensure proximity in the geographical space Fouedjio (2016), although this may be desirable if contiguous HRU are to be delineated.

The optimal number of clusters in the data is selected using the clustering tree and is calculated automatically by the function when the within variance reaches a minimum plateau, using the least number of clusters. In the method described by Husson et al. (2010), if $\Delta(Q)$ is the between inertia increase when moving from $Q-1$ to Q clusters, the optimal number of clusters Q is the one which minimize the relation $\Delta(Q)/\Delta(Q+1)$. Other indexes to assess the optimal number of clusters are described in Fouedjio (2016).

To test the present methodology, we calculate 10 scenarios in which each scenario has a number of s clusters (s from 1 to 10). For each scenario, the method stores for each cell the HRU it belongs to. This is done to evaluate the sensitivity of the hydrological model to the number of HRUs, ranging from a single HRU (a completely lumped model) to a more semi-distributed scheme of the basin with as many HRUs as clusters generated.

2.4 Hydrological model setup and simulations

The WEAP model is ran using the ten different scenarios described previously. Every configuration of the model uses a different number of HRU. The lumped configuration was called HRU_01 and uses just one HRU to model the basin. The second scenario uses two HRU and is called HRU_02. The rest of the configurations are called similarly depending on the number of HRU used.

The time series of monthly precipitation and mean monthly temperatures were derived from the meteorological dataset and the observed values recorded in the meteorological stations.

The values assigned for the hydrological parameters in each HRU are calculated as the average value for all the cells belonging to such HRU defined in the previous step. Also the values of Wind Speed, Relative Humidity and Albedo were set constant for every simulation for simplicity. These values were obtained by intersecting the area for each cluster defined in the previous section with the raster corresponding to the annual mean of each variable obtained from WRF outputs.

For calibration purposes, WEAP model has spatially-constant calibration factors for each of the four parameters assessed in the PCA/HCPC methodology. They are assumed initially as one but can be adjusted in the calibration process to adjust the results of the modelling by mean of automated or manual techniques.

Finally, the results of the hydrological modelling are analyzed by some common hydrological indicators as the Nash–Sutcliffe model efficiency coefficient (NSE) and the Root Mean Squared Error (RMSE) standardized by the mean discharge.

$$NSE = 1 - \frac{\sum_{i=1}^{N}(Q_{oi}-Q_{si})^2}{\sum_{i=1}^{N}(Q_{oi}-\overline{Q_o})^2}, \qquad (2)$$

$$RMSE = \frac{1}{\overline{Q_o}} \cdot \sqrt[2]{\frac{\sum_{i=1}^{N}(Q_{oi}-Q_{si})^2}{N}}, \qquad (3)$$

Where:

Q_{oi}: Observed discharge at time step i.

Q_{si}: Simulated discharge at time step i.

$\overline{Q_o}$: Average of the observed discharge over the simulation period.

3 Results

This section shows the main results in each part of the methodology: (i) the results of the Principal Component Analysis and the Hierarchical Clustering and (ii) the results of the hydrological modelling using the different schemes of the HRU.

3.1 PCA and cluster analysis

This section shows the results of the Principal Component Analysis and the Hierarchical Clustering Analysis, the core of the HRU delineation. The PCA was performed over the set of the meteorological variables (Precipitation, Temperature, Relative Humidity, Wind Velocity, Albedo and Evapotranspiration) and hydrologic parameters (*Sw*, *f*, *RRF* and *Ks*). The first result to highlight is that the two first dimensions resulting from the PCA account for 66.8% of the total variance. Adding the following 3rd, 4th and 5th dimensions, they account for the 78.7%, 84.6% and 89.6% respectively of the total variance of the master dataset. Table 1 shows the variance explained by each consecutive eigenvector or dimension and the contribution of each variable to the dimensions. The first dimension (more than 50% of the total variance) is composed mainly on meteorological variables, being Temperature, Albedo, Wind Speed and Rainfall the ones with more contribution. The second-dimension accounts for more than 16.1% of the total variance and is composed mainly by rainfall and the hydrological parameters.

Table 1. Summary from the PCA analysis results for the first five dimension. Where the upper part shows the total variance explained by each dimension and lower, the contribution of each variable to that dimension.

	Dim.1	Dim.2	Dim.3	Dim.4	Dim.5
Variance Explained (%)	50.7	16.1	11.9	5.9	5.0
Variables	Contribution to each dimension (%)				
Temp	28.2	0.1	0.7	0.2	5.1
Albedo	11.5	1.1	5.6	1.6	10.0
WS	10.8	1.3	0.6	0.0	8.5
Pp	10.7	41.5	22.6	0.0	0.6
EVPM	10.2	2.2	1.5	0.5	12.3
RNet	9.7	3.5	9.6	1.4	10.7
HR	9.5	5.6	1.4	1.0	0.0
RRF	4.9	13.9	10.5	12.1	2.5
Sw	4.1	20.3	6.4	4.4	9.7
F	0.3	4.7	32.1	0.4	35.6
Ks	0.0	5.7	9.0	78.3	5.0

Table 1 suggests that the firsts five dimensions are carrying most of the information, cleaning the statistical noise and hence these first five components will be used in the HCPC function for the cluster analysis. Based on the new dataset composed by only these principal components, the total variance of the system is fixed for the cluster analysis, therefore the within variance will be expressed as relative to such total hereinafter.

Figure 4 shows the proportion of the within variance relative to the total variance, where the major decrement of the variance occurs up to the case with four clusters and decreases until the case with seven or eight clusters. A decrease in the within variance means that the internal variability of each cluster decreases and hence the variance between clusters increases (Equation 1). As this happens, the hydrologic behavior between HRUs is also expected to be more heterogeneous and simultaneously more homogeneous within each individual HRU, which is expected to lead in a better hydrologic modelling. Hence, it was expected that the optimal number of HRU for hydrological modeling is four. As described in the methodology, the Hierarchical Cluster Analysis was performed to produce ten scenarios with different number of HRUs partitioning the basin, varying between one (lumped model; HRU_1) to ten (HRU_10), to be tested in the hydrological model.

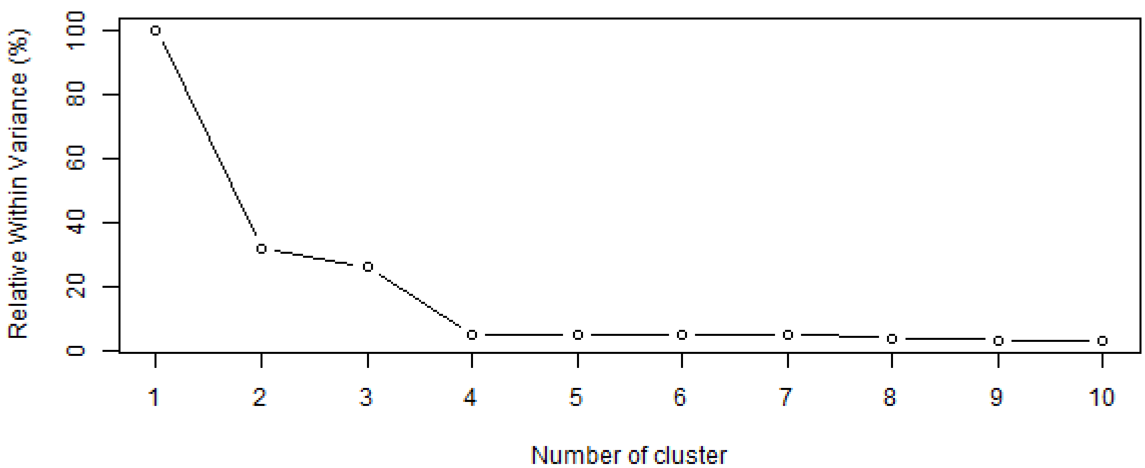

Figure 4. Relative within variance for each number of clusters.

Figure 5 shows the distribution of cells in six equal elevation bands, roughly of 550m each (left plot), as generally used in WEAP as first step for separation of HRUs (Vicuña et al., 2010; Young et al., 2009) and six clusters following the methodology proposed in this work, named PCA/HCPC (right). This number of clusters was chosen because of the best hydrologic results (section 3.3). Cluster 6 is similar in shape with the highest elevation band as they are concentrated in the eastern part of the basin were the highest elevations are located. For other clusters, the figure shows a clear difference; for instance, cluster 2 in the PCA/HCPC methodology is concentrated in the northwestern part of the catchment, consistent with the high precipitation area identified in Figure 3a, which is not identified in the traditional elevation bands. The cluster with the lowest elevations (cluster 1) is not as regular as its corresponding elevation band, as this cluster seems to follow the riverbed and the flat riparian zone. Cluster 5 seems to be concentrated in higher and colder areas with Andean vegetation and vegas, characterized by their high-water content or retention capacity, compared to the surroundings composed mainly by bare soil and disperse and small shrubs (cluster 4). Clusters still follow a tendency by elevation, as Mean Annual Temperature is the main variable composing the first dimension, but other variables tend to get importance as the number of clusters increase.

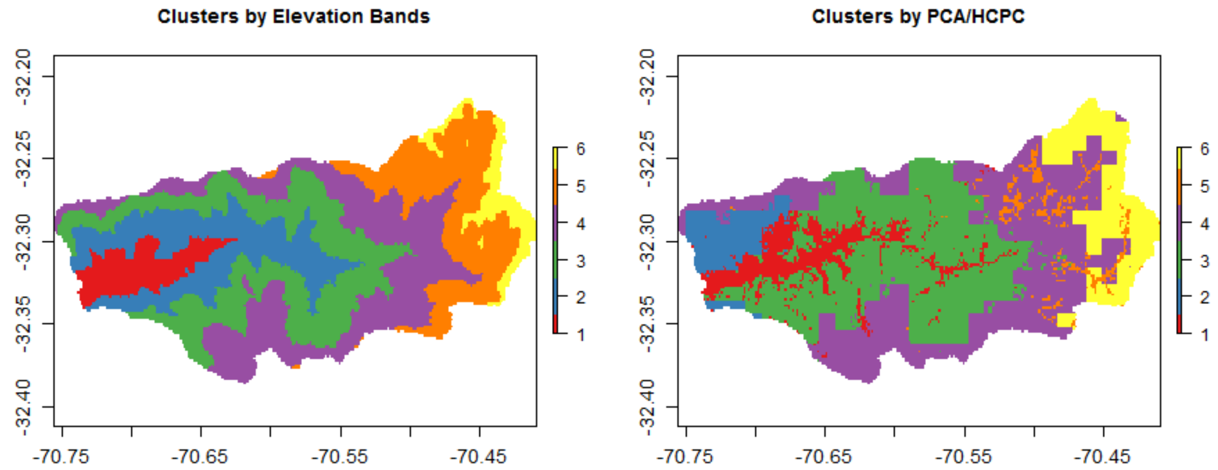

Figure 5. Elevations bands every 550m (left), as used in the traditional methodology, and HRU delimitations (right) for the simulation with six HRUs by the PCA/HCPC methodology.

Figure 6 shows the boxplot for the values of six selected variables on each cell grouped by cluster to highlight the differences between them. It is possible to observe distinct characteristics between clusters: cluster 6 is the coldest cluster and has one of the highest precipitation rates; cluster 1 is the warmest, the lowest in altitude, very dry and the one having the deepest soil capacity, probably due to its location in the deepest part of the valley, where soils tend to be deeper and with higher runoff resistance factor, due to its flat terrain and vegetation. Cluster 2 is the one with highest precipitation. Cluster 5 is the one with highest hydraulic conductivity, due to the presence of marshes and wetlands. Clusters 3 and 4 are similar although cluster 4 has a mean value for precipitation of almost 50% more than cluster 3, and also have differences in the variables not shown. Clusters 6 and 4 have similar hydrologic parameters, but cluster 6 is colder and receives more rainfall.

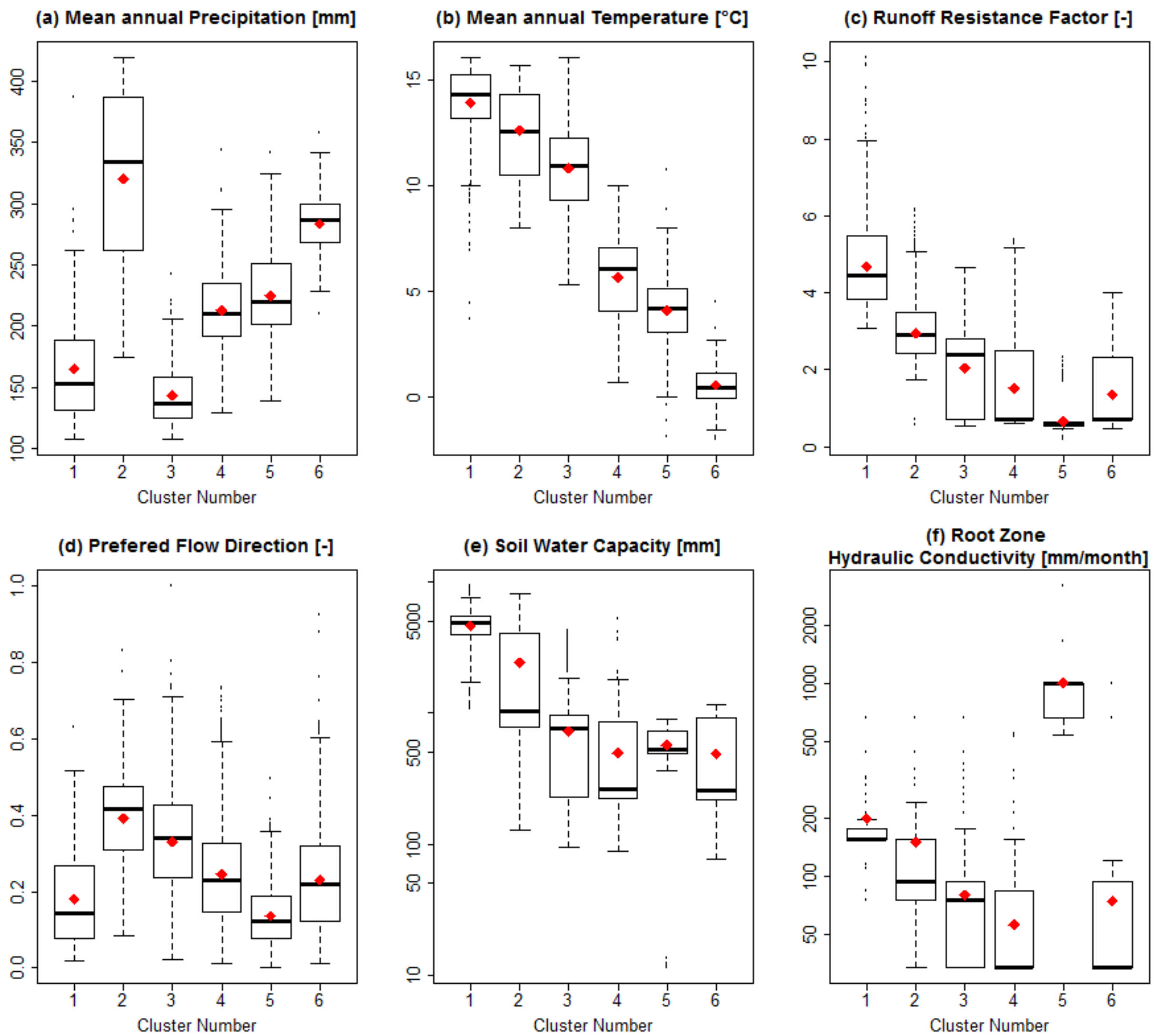

Figure 6. Boxplots per cluster in the simulation HRU_06 for: (a) Mean annual precipitation, (b) Mean temperature, (c) RRF, (d) f, (e) Sw and (f) Ks. Red circles represent the mean value of each cluster. Sw and Ks are plotted in log scale. Means are shown as red dots.

3.2 Hydrological Modelling and HRUs contribution

The WEAP Model was ran using ten scenarios, each one with a different number of HRUs ranging from 1 to 10 (labeled as HRU_01 to HRU_10). Each simulation was run in a monthly time step starting from April 1979 until March 2016, following the Water Year commonly used in Chile, although the first four years were dismissed due to the warming of the model.

As the number of HRU increases and the level of spatial discretization is more detailed, also the model efficiency increases. For the first simulation with a lumped scheme, the NSE and RMSE (Equation 2 and Equation 3) values were 0.58 and 4.1% respectively and both indexes improved as more HRU were used. But for simulations with more than six HRUs, the extra clusters or HRU are not making any considerable improvement in the results, consistent with what was described in Haverkamp et al. (2005) and the model efficiency fluctuates in a plateau of 0.76-

0.79 while the RMSE near 3.1-3.2%. Figure 7 shows the observed and simulated monthly hydrograph for HRU_06.

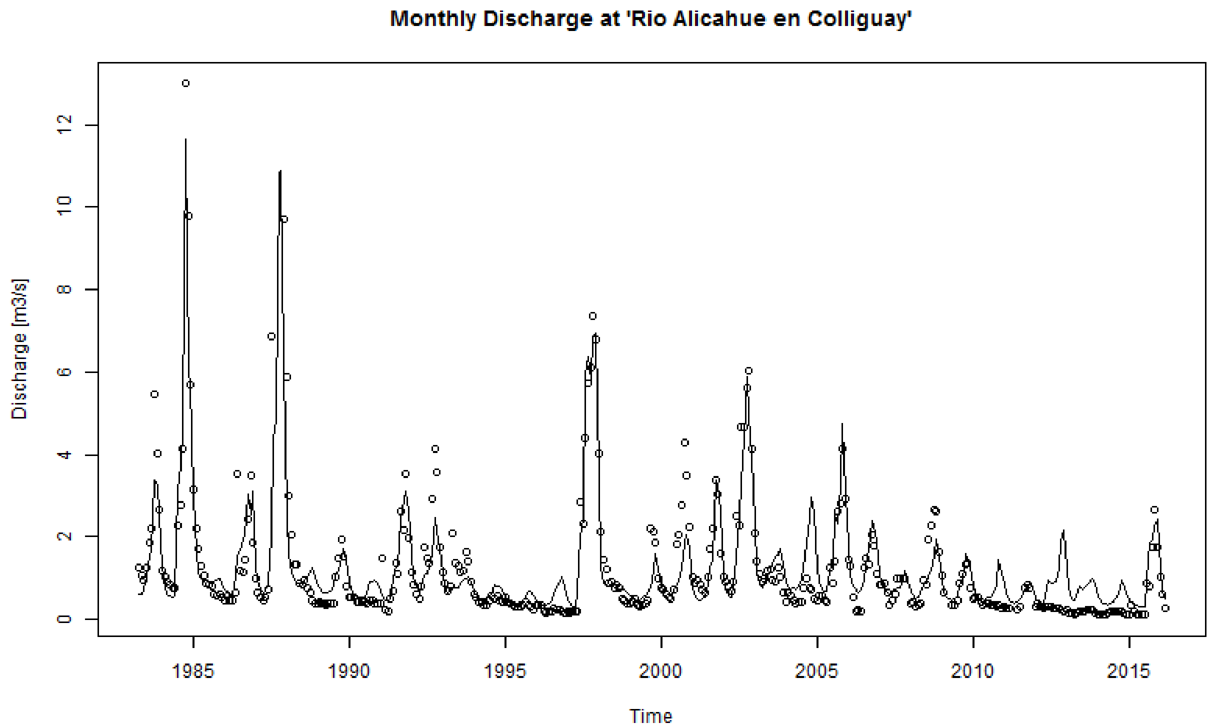

Figure 7. Hydrograph of observed and modeled streamflow in 'Rio Alicahue en Colliguay' station for the simulation HRU_06. Observed discharge is shown in dots and the simulated discharge in a continuous line.

Table 2 presents the differences between clusters in terms of inputs and responses and it is used to assess the different hydrological processes that each HRU represent. It shows the mean annual temperature, rainfall and elevation, the mean annual discharge and its standard deviation, the variation coefficient and the centroid of the annual flow volume as an index to measures timing of peak discharge in the season, calculated as a weighted average of the month and the discharge associated to each month (Young et al., 2009):

$$Hydro_{Centroid} = \frac{\sum_{i=1}^{N} month\_index_i \cdot Q_{si}}{\sum_{i=1}^{N} Q_{si}}, \qquad (4)$$

, where January has index 1 and December, 12.

Although clusters 6 and 3 have similarities in total discharge, its annual variation, cluster 6 groups most of the coldest cells in the basin where the snow melts late during the season and its discharge center of mass is in the middle of November and peaks in January, while cluster 3 peaks in the middle of September, coinciding with its center of mass of the hydrograph. Both clusters are controlled by very different hydrological process and parameters.

Interest is on cluster 5, as it has a relatively small amount of area but its proportional contribution to the total discharge doubles its relative area. It peaks at the beginning of summer, has the lower Evaporation/Precipitation ratio and it is an example of an extreme hydrologic behavior that must be characterized and not dismissed. It is possible to argue that clusters 5 and 4 can be merged, as their peak at the same time and they have the same elevation, but that decision depends on to what extent it is possible to aggregate.

Clusters 1 and 2 are the lower in elevation, but the relative contribution of cluster 1 compared to its relative area indicates that is the less important and its discharge is comparable in absolute terms to cluster 2, although their areas are 34.0 and 22.1 km^2.

Table 2. Summary of mean annual variables for the six clusters used in the hydrological simulation. The values correspond to the mean values for the simulation period of 1984-2016.

	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5	Cluster 6
Area [Km^2]	34.0	22.1	125.2	121.7	10.9	41.5
% over total area	9.6%	6.2%	35.2%	34.2%	3.1%	11.7%
Elevation [m]	1483	1460	2063	2080	2837	2951
Precipitation [mm]	204	413	175	269	287	357
Evapotranspiration [mm]	126	203	110	154	50	161
E/P [-]	0.62	0.49	0.63	0.57	0.17	0.45
Discharge						
***Mean* [m3/s]**	0.09	0.15	0.26	0.45	0.08	0.26
% over total discharge	7%	12%	20%	35%	6%	20%
***Standard Deviation* [m3/s]**	0.03	0.11	0.30	0.77	0.08	0.47
Coefficient of Variation	0.34	0.76	1.16	1.72	0.99	1.79
Hydrograph Centroid [month index]	9.65	9.85	9.79	10.57	10.35	11.49

Finally, Figure 8 presents the mean monthly discharge (left) and the mean monthly areal discharge production (right) for each of the six clusters. Each cluster shows different hydrographs in terms of total volume and peak timing, consistent with the goal of maximization of the between-variability; this behavior is also seen in other scenarios with different number of clusters, although not shown here. Cluster 4 is the main contributor to the annual discharge, and it is clear its nival hydrologic regime with its peak in November (half Spring in Southern Hemisphere), coinciding with the basin peak due to snowmelt. Clusters 1, 2 and 3 present hydrographs with peaks in or near September, two months later of the precipitation peaks for this region during austral winter, probably due to the firsts snowmelts but also from interflow produced from rainfall that reacts slower than direct runoff. Cluster 1 is also the more stable in terms of discharge, mainly due to the availability to hold water because of its larger soil water capacity and cluster 3 presents a more distinct peak in the end of the winter probably due to the first snowmelt but also rainfall and humidity leaving the upper part of the soil. Cluster 6 presents the most retarded hydrograph peak in the season, explained by late melt of snow due to its relatively higher mean elevation compared to the other HRUs, hence, the lowest values of mean temperatures. Clusters 4 and 5 also present a nival regime as their peaks match with the snow melting season, but its total volume is completely different as cluster 5 is explained by a concentration in a relatively small area of marshes and Andes wetlands while cluster 4 shows the biggest contribution to the total streamflow, mainly given by its high portion of area (34.2%).

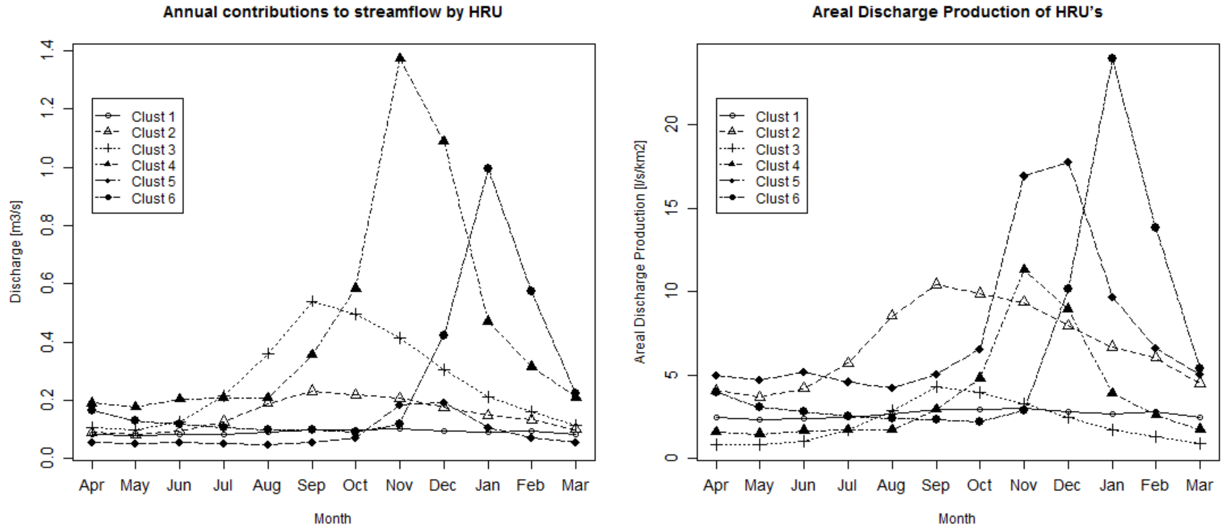

Figure 8. Mean monthly discharge of each HRU contributing to total streamflow at the outlet, for the six HRUs scenario. Left panel: mean annual discharge. Right panel: Mean annual discharge production by area.

The image at the right shows the production of discharge relative to the area of the HRU (in l/s/km^2). The relative importance of each HRU changes, especially for clusters 2, 5 and 6. As shown in Table 2, the relative contribution to the total discharge of those clusters doubles their area relative to the total area. The hydrologic regime of cluster 2 tends to be closer to the precipitation season (May to August) and it has a high areal production of water due mainly to the concentration of rainfall in that area of the catchment. Cluster 5 presents the higher average of areal production (7.6 l/s/km^2) and even its base value of near 5.0 l/s/km^2 is also higher than the rest of the base values. Again, this may be explained by the nature of the vegetation covering most of that cluster and by the slow release of water stored in them. The highest peak of 23.9 l/s/km^2 in the month of January corresponds to cluster 6 and it is a combination of high rates of precipitation during winter in a relatively small area of the catchment, accumulating a massive volume of snow with the rise of temperatures in Summer, producing the highest peak of discharge per unit of area due to snowmelt.

4 Discussion

This paper presents a new methodology for HRU delineation based on the catchment attributes, explained by the model parameters and climate variables. The units generated are expected to be used in lumped and semi-distributed hydrological models where the topology of the elements could be neglected. The methodology present two main steps: (i) a Principal Component Analysis to reduce the number of variables while most variance is kept, and (ii) a Hierarchical Clustering decomposition to delineate the HRUs with minimum internal variability but maximum variability among the created units.

The methodology was tested on the Alicahue River Basin with the WEAP model which has a hydrology module. The model was run under ten scenarios with different numbers of clusters (HRUs) and evaluated using the Nash-Sutcliff Efficiency Index and the Mean Squared Error.

4.1 Methodology and data uncertainties in the dataset preparation

Although the generation of the complete dataset used to derive the HRUs was not presented, the estimation of the parameters and the climate dataset can be calculated independently using any methodology or information previously available. The only two main characteristics that must be preserved from such methodology are: (1) climate information needs to be from a gridded dataset and (2) hydrologic parameters must be specific to the target model and calculated spatially prior to the HRU delineation.

The main reason to use a WRF simulation, instead of longer and publicly available datasets, was its high resolution (1 km) which is very important in regions with very complex topography as the Alicahue basin. That resolution is much higher than Reanalysis (Kalnay et al., 1996) which is usually about 0.5º and even higher than the newest local dataset CR2MET available at 0.05° for the continental Chilean territory (Alvarez-Garreton et al., 2018; DGA, 2017). But it is important to highlight the evident limitation on the case of study due to have only 3 years of WRF simulations to represent the climate of the region.

4.2 Clustering method and Results

Once the data was pre-processed, it contained more than 17 thousand cells and each cell had information on four hydrologic parameters and seven meteorological variables; moreover, in other implementations this number could be even bigger. That amount of data had to be summarized in order to be treatable, but at the same time, it was desirable that the aggregation process carried most of the variability, without losing valuable information. The Principal Component Analysis was chosen, as it selects orthogonal vectors whom carry much of the information gathered in the previous process.

The PCA function uses weights to account for the relative importance of the variables. This gives to the modeler the option to assess the most important variables given the model and/or the problem to solve. In this study case, weights for rainfall and temperature were equal to 2 while for all other variables it was set to 1. Temperature plays a crucial role controlling evapotranspiration and snow melting, both main hydrological characteristics of an Andean semi-arid basin, as the Alicahue basin in the study case and Precipitation controls the water income to the basins and water simulations are highly sensitive to the amount of water used to model. Sensitivity analysis of PCA to the weights was not performed, but it could help gain information about the robustness of the clusters given different weights.

The within variance decrease obtained from the clustering process could be used as a criterion for the selection of the optimal number of HRUs required to capture the main hydrological behaviors in the target basin. This methodology allows to highly reduce the number of HRUs involved in the simulation, decreasing the required computational time. This could favor, for instance, studies with ensemble simulations, more exhaustive sensitivity analysis to some parameters of the models and/or much longer (or higher temporal resolutions) simulations.

4.3 Discharge independence in the hydrologic modelling

The results by the hydrological modelling show in general a good match between observations and simulations, even with the lumped scheme. As expected, the simulation with one HRU, as the most lumped scheme, show the poorest results in terms of efficiency (NS=0.58). As the number of HRU increases and the level of spatial discretization is more detailed, also the model efficiency increases. The errors in the modelling can be explained, at least to a good extent, by the uncertainties in the simple meteorological models used to derive the precipitation and the temperature, the proposed relations to obtain the parameters maps and a possible lack of representing all the hydrologic fluxes and storages in the hydrologic cycle in the basin given by lack of soil information. Also the representation of extreme hydrologic phenomena is possible only if the chosen model is capable of simulate these phenomena. If not, any discretization of HRU methodology would be useless or at least not useful.

The streamflow at the outlet of Alicahue basin is controlled by a baseflow dominated by the subsurface storage which is dependent in the storage capacity of soil and evapotranspiration stress, a component driven by the winter rainfall dependent in hydraulic conductivity and rain intensity, and a component driven by snow melting which is highly dependent in temperature and elevation. From Figure 8, such behaviors were well captured for the different HRUs, which allows to the modeler a better understanding of the underlying processes controlling the outlet streamflow when compared to other methodologies for HRU delimitation.

Finally, it is important to note that the only variable for assessing the methodology was the river discharge, which simplify the water cycle and all its components into one lumped criterion. It would be advisable test the methodology and the hydrologic behavior with the rest of the components of the hydrological cycle (infiltration, evapotranspiration, groundwater movement, leakage, etc.), which was not possible in the case of study basin due to the lack of observations.

5 Conclusions

Flügel presented in 1983 the concept of HRU for the hydrological modelling. HRUs are the basic units in which the equations controlled by parameters are run and meteorological data is used as inputs. The basic assumptions of HRU is that each of them has a particular hydrological response to rainfall, temperature and other climate data. Most of the actual methodologies account only partially for the spatial variability that leads to differentiated response, particularly the spatial climate variability within the basin is under- or misrepresented.

This paper presented a methodology for the determination of HRU, more consistent with the classical definition, based on hydrological parameters (specifics to the target model) and meteorological inputs; using Principal Component Analysis and Hierarchical Clustering to minimize the global internal variability in each HRU and that at the same time maximizes the variability among HRUs. This procedure is intended to generate different responses by each unit, as defined by the modeled hydrograph, minimizing the number of required HRUs to capture the internal variability.

The application of the methodology was assessed in the Alicahue river basin, a small basin located in a semi-arid and mountainous region in Central Chile, with altitudes ranging from the 780 to almost 4000 meters above sea level. Results of the WEAP simulations shows a good agreement between modeled and observed streamflow at the outlet, with scores comparable to other studies using the same model in similar basins.

Better hydrological parameters and meteorological datasets could still improve the model efficiency. Future research is to test the methodology in other basins with different hydrologic regime and using different models. WEAP is suitable for time steps longer than one day, but the methodology can be used in other long-term models or even in storm models, considering other parameters sensible to the basin response (i.e. concentration time, curve number, etc.).

In summary, the main advantages of the proposed methodology are:

Computational efficiency in the hydrological simulations. As the methodology is designed to minimize the required numbers of HRU to account for most of the spatial variability in the climate and hydrological parameters (the main controllers of the hydrological response), the computational effort is highly reduced as usually it is linear in the number of HRUs. In the study case only six HRUs were necessary to achieve similar scores than those from more commons methodologies that use several tens of HRUs.

Basin heterogeneity better captured. As the PCA captures most of the variability of the parameters and climate variables, heterogeneous conditions are kept even after the reduction of the number of variables used as input to the cluster analysis. Also, the hierarchical clustering process ensure the delineation of the HRUs is completely driven by such variability and not by arbitrary choices. For instance, in the study case, one of the HRUs correspond to a small and disjoint area that has a relatively large contribution to the total streamflow, which would be probably neglected with most of the traditional methodologies.

Better identifiability of the HRUs. As the HRU delineation was driven by the minimization of the within variance in each HRU and at the same time maximization of the variance between HRUs, the hydrological response is expected to be different for each HRU with minimum redundancy. This will allow to the modeler to gain a better understanding of the underlying hydrological behaviors that controls the response of the basin. For instance, each HRU in the case of study was identified with different processes, including baseflow, quick rainfall-runoff response, snow melting at different times associated with elevation and temperature differences.

6 Acknowledgments, Samples, and Data

Supplementary data is available at DOI: 10.17632/ppgtgvyttm.2.

7 References

Alvarez-Garreton, C., Mendoza, P. A., Pablo Boisier, J., Addor, N., Galleguillos, M., Zambrano-Bigiarini, M., … Ayala, A. (2018). The CAMELS-CL dataset: catchment attributes and meteorology for large sample studies-Chile dataset. *Hydrol. Earth Syst. Sci*, *22*, 5817–5846. https://doi.org/10.5194/hess-22-5817-2018

Bonelli, S., Vicuña, S., Meza, F. J., Gironás, J., & Barton, J. (2014). Incorporating climate change adaptation strategies in urban water supply planning: The case of central Chile. *Journal of Water and Climate Change*, *5*(3), 357–376. https://doi.org/10.2166/wcc.2014.037

Dehotin, J., & Braud, I. (2008). Which spatial discretization for distributed hydrological models? Proposition of a methodology and illustration for medium to large-scale catchments. *Hydrology and Earth System Sciences*, *12*(3), 769–796. https://doi.org/10.5194/hess-12-769-2008

DGA. (2017). *Actualización del Balance Hídrico Nacional, SIT N° 417.* Santiago. Retrieved from http://documentos.dga.cl/REH5796v1.pdf

Flügel, W.-A. (1995). Delineating hydrological response units by geographical information system analyses for regional hydrological modelling using PRMS/MMS in the drainage basin of the River Bröl, Germany. *Hydrological Processes*, *9*(3–4), 423–436. https://doi.org/10.1002/hyp.3360090313

Fouedjio, F. (2016). A hierarchical clustering method for multivariate geostatistical data. *Spatial Statistics*, *18*, 333–351. https://doi.org/10.1016/j.spasta.2016.07.003

Garreaud, R. (1992). Impact of the variability of snowline in winter discharge peaks in basins of mixed regime in Central Chile (in spanish). Sociedad Chilena de Ingeniería Hidráulica.

Haghnegahdar, A., Tolson, B. A., Craig, J. R., & Paya, K. T. (2015). Assessing the performance of a semi-distributed hydrological model under various watershed discretization schemes. *Hydrological Processes*, *29*(18), 4018–4031. https://doi.org/10.1002/hyp.10550

Han, J.-C., Huang, G.-H., Zhang, H., Li, Z., & Li, Y.-P. (2014). Effects of watershed subdivision level on semi-distributed hydrological simulations: case study of the SLURP model applied to the Xiangxi River watershed, China. *Hydrological Sciences Journal*, *59*(1), 108–125. https://doi.org/10.1080/02626667.2013.854368

Haverkamp, S., Fohrer, N., & Frede, H. G. (2005). Assessment of the effect of land use patterns on hydrologic landscape functions: A comprehensive GIS-based tool to minimize model uncertainty resulting from spatial aggregation. *Hydrological Processes*, *19*(3), 715–727. https://doi.org/10.1002/hyp.5626

Hijmans, R. J. (2017). Geosphere: Spherical Trigonometry. R package version 1.5-7. https://CRAN.R-project.org/package=geosphere.

Höge, M., Wöhling, T., & Nowak, W. (2018). A Primer for Model Selection: The Decisive Role of Model Complexity. *Water Resources Research*, *54*(3), 1688–1715. https://doi.org/10.1002/2017WR021902

Husson, F., Julie, J., & Jérôme, P. (2010). Technical Report – Agrocampus Principal component methods -hierarchical clustering -partitional clustering: why would we need to choose for visualizing data? *Applied Mathematics Department Agrocampus*.

Kalnay, E., Kanamitsu, M., Kistler, R., Collins, W., Deaven, D., Gandin, L., … Joseph, D. (1996). The NCEP/NCAR 40-Year Reanalysis Project. *Bulletin of the American Meteorological Society*, *77*(3), 437–471. https://doi.org/10.1175/1520-0477(1996)077<0437:TNYRP>2.0.CO;2

Lê, S., Josse, J., & Husson, F. (2008). FactoMineR: An R Package for Multivariate Analysis. *Journal of Statistical Software*, *25*(1).

Leavesley, G. H., Lichty, R. W., Troutman, B. M., & Saindon, L. G. (1983). Precipitation-runoff modeling system; user's manual. https://doi.org/10.3133/wri834238

Martínez, E., Flores, J., Retamal, M., Ahumada, I., & Brito, S. (2013). *Informe técnico final monitoreo de cambios, corrección cartográfica y actualización del catastro de bosque nativo en las Regiones de Valparaíso, Metropolitana y Libertador Bernardo O'Higgins*.

Nijzink, R., Hutton, C., Pechlivanidis, I., Capell, R., Arheimer, B., Freer, J., … Hrachowitz, M. (2016). The evolution of root-zone moisture capacities after deforestation: a step towards hydrological predictions under change? *Hydrol. Earth Syst. Sci*, *20*, 4775–4799. https://doi.org/10.5194/hess-20-4775-2016

Orth, R., Dutra, E., & Pappenberger, F. (2016). Improving weather predictability by including land surface model parameter uncertainty. *Monthly Weather Review*, *144*(4), 1551–1569. https://doi.org/10.1175/MWR-D-15-0283.1

Pilz, T., Francke, T., & Bronstert, A. (2017). LumpR 2.0.0: An R package facilitating landscape discretisation for hillslope-based hydrological models. *Geoscientific Model Development*, *10*(8), 3001–3023. https://doi.org/10.5194/gmd-10-3001-2017

Purkey, D. R., Joyce, B., Vicuña, S., Hanemann, M. W., Dale, L. L., Yates, D., & Dracup, J. a. (2007). Robust analysis of future climate change impacts on water for agriculture and other sectors: a case study in the Sacramento Valley. *Climatic Change*, *87*(S1), 109–122. https://doi.org/10.1007/s10584-007-9375-8

R Core Team. (2019). R: A Language and Environment for Statistical Computing. Vienna, Austria. Retrieved from https://www.r-project.org/

Savvidou, E., Efstratiadis, A., Koussis, A. D., Koukouvinos, A., & Skarlatos, D. (2018). The curve number concept as a driver for delineating Hydrological Response Units. *Water (Switzerland)*, *10*(2). https://doi.org/10.3390/w10020194

Skamarock, W. C., Klemp, J. B., Dudhia, J., Gill, D. O., Barker, D. M., Duda, M. G., … Powers, J. G. (2008). *A Description of the Advanced Research WRF Version 3*. Retrieved from http://opensky.ucar.edu/islandora/object/technotes%3A500/datastream/PDF/view

Tachikawa, T., Kaku, M., Iwasaki, A., Gesch, D. B., Oimoen, M. J., Zhang, Z., … Carabajal, C. (2011). ASTER global digital elevation model version 2 - summary of validaton results. NASA Land Processes Distributed Active Archive Center. Downloaded at: http://www.jspacesystems.or.jp/ersdac/GDEM/ver2Validation/Summary_GDEM2_validation_report_final.pdf, (March 2015).

Vicuña, S., Garreaud, R. D., & McPhee, J. (2010). Climate change impacts on the hydrology of a snowmelt driven basin in semiarid Chile. *Climatic Change*, *105*(3–4), 469–488. https://doi.org/10.1007/s10584-010-9888-4

Yates, D., Sieber, J., Purkey, D., & Huber-Lee, A. (2005). 21 - A Demand-, Priority-, and Preference-Driven Water Planning Model. Part 1: Model Characteristics. Water International 30, 487-500., *30*(4), 487–500.

Young, C. A., Escobar-Arias, M. I., Fernandes, M., Joyce, B., Kiparsky, M., Mount, J. F., … Yates, D. (2009). Modeling the hydrology of climate change in California’s Sierra Nevada for subwatershed scale adaptation. *Journal of the American Water Resources Association (JAWRA)*, *45*(6), 1409–1423. https://doi.org/10.1111