ECOHYDROLOGY
Ecohydrol. (2016)
Published online in Wiley Online Library
(wileyonlinelibrary.com) DOI: 10.1002/eco.1721

Towards catchment classification in data-scarce regions

Daniel A. Auerbach,[1]* Brian P. Buchanan,[2] Alexander V. Alexiades,[3] Elizabeth P. Anderson,[4] Andrea C. Encalada,[5] Erin I. Larson,[1] Ryan A. McManamay,[6] Gregory L. Poe,[7] M. Todd Walter[2] and Alexander S. Flecker[1]

[1] *Department of Ecology and Evolutionary Biology, Cornell University, Ithaca, NY, USA*
[2] *Department of Biological and Environmental Engineering, Cornell University, Ithaca, NY, USA*
[3] *Department of Natural Resources, Cornell University, Ithaca, NY, USA*
[4] *School of Environment, Arts and Society, Florida International University, Miami, FL, USA*
[5] *Laboratorio de Ecología Acuática, Colegio de Ciencias Biológicas y Ambientales, Universidad San Francisco de Quito, Quito, Ecuador*
[6] *Environmental Sciences Division, Oak Ridge National Lab, Oak Ridge, TN, USA*
[7] *Dyson School of Applied Economics and Management, Cornell University, Ithaca, NY, USA*

ABSTRACT

Assessing spatial variation in hydrologic processes can help to inform freshwater management and advance ecological understanding, yet many areas lack sufficient flow records on which to base classifications. Seeking to address this challenge, we apply concepts developed in data-rich settings to public, global data in order to demonstrate a broadly replicable approach to characterizing hydrologic variation. The proposed approach groups the basins associated with reaches in a river network according to key environmental drivers of hydrologic conditions. This initial study examines Colorado (USA), where long-term streamflow records permit comparison with previously distinguished flow regime types, and Ecuador, where data limitations preclude such analysis. The flow regime types assigned to gages in Colorado corresponded reasonably well to the classes distinguished from environmental features. The divisions in Ecuador reflected major known biophysical gradients while also providing a higher resolution supplement to an existing depiction of freshwater ecoregions. Although freshwater policy and management decisions occur amidst uncertainty and imperfect knowledge, this classification framework offers a rigorous and transferrable means to distinguish catchments in data-scarce regions. The maps and attributes of the resulting ecohydrologic classes offer a departure point for additional study and data collection programmes such as the placement of stations in under-monitored classes, and the divisions may serve as a preliminary template with which to structure conservation efforts such as environmental flow assessments.

Supporting information may be found in the online version of this article.

KEY WORDS environmental classification; hydroclimate; flow regime; Colorado; Ecuador

Received 23 March 2015; Revised 16 December 2015; Accepted 17 December 2015

INTRODUCTION

Managers worldwide are challenged to make informed decisions about dam construction, river diversion, land conversion and other development actions that affect freshwater biodiversity and resource allocation. The challenge is especially acute in settings where rapidly growing human populations are vulnerable to flooding and inadequate drinking water supplies (Vorosmarty *et al.*, 2010). A data-driven description of the spatial variation in hydrologic processes is therefore critical to understanding water resource availability and to defining freshwater conservation guidelines. Yet the data, time and money available for such assessments are scarce globally, particularly in tropical areas with abundant biodiversity (Abell *et al.*, 2008). In this context, a classification system can aid research and management by translating the full spectrum of natural variation into a tractable set of groups.

Classifications based on statistical similarity in the long-term discharge records at gages in relatively unimpaired basins can define natural or 'normative' flow types where such data exist (Olden *et al.*, 2012; Archfield *et al.*, 2014). Such approaches, termed 'inductive' by Olden *et al.* (2012), follow increasingly well-established protocols for which measures to calculate and how to use them to aggregate groups. These classifications also sometimes incorporate geologic, chemical or temperature information, and research

*Correspondence to: Daniel A. Auerbach, Department of Ecology and Evolutionary Biology, Cornell University, Ithaca, NY, USA.
E-mail: da373@cornell.edu
[#]ORISE Fellow, Office of Wetlands, Oceans and Watersheds, US Environmental Protection Agency, Washington DC
[#]Heritage University, Department of Environmental Sciences, Toppenish, WA

has begun to address the quantification of uncertainty (Reidy Liermann *et al.*, 2012b; McManamay *et al.*, 2014). Yet, the necessary streamflow data are unavailable or insufficient in many regions, many of which are also those under the greatest pressure to develop water use infrastructure (Vorosmarty *et al.*, 2010; Richter *et al.*, 2011). In such settings, a 'deductive' classification therefore offers an important means to capture key differences by basing classes on the more readily available geographic environmental data that reflect the main drivers of hydrologic processes (Olden *et al.*, 2012; Wagener *et al.*, 2007).

Numerous studies have demonstrated the utility of deductive approaches in areas with relatively abundant hydrological and biological data. For example, Snelder and Biggs (2002) established a conceptual framework for river network classification that specified the dominant processes acting across a hierarchy of spatial scales: climate, source of flow (surface versus groundwater), geology, land cover, network position and valley form. When applied to New Zealand (Snelder and Biggs, 2002; Snelder *et al.*, 2005), France (Snelder *et al.*, 2009) and Spain (Peñas *et al.*, 2014), this framework has successfully generated information to support research and management. In the United States, Wolock *et al.* (2004) identified 20 non-contiguous 'hydrologic landscape regions' (HLRs) by clustering basins according to similar climate, topographic and geologic attributes (grouping 43 931 units, each approximately 200 km^2). These HLRs effectively predicted environmental variance in catchments above a set of nationwide water quality sampling sites, although prediction strength was lower for in-channel nutrient conditions and fish species richness. Higgins *et al.* (2005) also illustrated the use of four tiered spatial scales to distinguish classes based on combinations of physical and biological data, such as dominant geology, historical distributions of native fish and patterns of watershed connectivity. Application of this method in the basins of the Columbia and Upper Paraguay Rivers underscored the value of a classification system to freshwater conservation planning and revealed challenges for implementing such analyses in data-scarce locations. Finally, Sawicz *et al.* (2011) clustered 280 undeveloped basins throughout the eastern US according to physical attributes such as mean elevation, percentage agricultural land and number of days with precipitation. They found that class membership effectively represented differences in six hydrologic signatures relevant to management objectives (e.g. the runoff ratio and the shape of the flow duration curve). Recent research in data-rich areas continues to advance our understanding of best practices for streamflow classification and how catchment similarity corresponds to streamflow properties (Archfield and Vogel, 2010; Snelder and Booker, 2012; Archfield *et al.*, 2014; Sawicz *et al.*, 2014).

Such studies have clarified many important issues for hydrological classification, but additional work is needed to extend the ideas further into data-scarce settings using methods that are applicable both globally and at relatively fine spatial scales. Higgins *et al.* (2005) identified circumstances in which limited data might necessitate a 'top-down' approach, drawing more heavily on expert opinion and disaggregating focal units rather than aggregating up from fine-grained measures. Nonetheless, the increasing availability of medium-resolution and high-resolution biophysical data sets, alongside increases in computing power and advances in analysis software, has created the potential to define 'bottom up' classifications built by clustering spatial units according to data. These classifications could refine understanding of broad hydrologic patterns across watersheds and reveal finer scale patterns relevant to organisms and ecosystem functions.

Accordingly, our research objective was to develop a broadly replicable catchment classification approach that could be applied in areas with little to no streamflow data. Building on the concepts established by Poff (1997), Snelder and Biggs (2002), Wolock *et al.* (2004), Higgins *et al.* (2005) and others, we analyzed several readily accessible global data sets on topography, climate, soils and land cover in conjunction with catchments delineated at multiple spatial scales worldwide. We compare the proposed approach with a recent inductive classification based on high-quality streamflow data in Colorado (McManamay *et al.*, 2014) and apply the method to the mainland portion of Ecuador as a data-scarce case study, thereby providing an initial test of performance in two steep, biophysically diverse landscapes.

METHODS

Study areas

Ecuador and Colorado encompass similarly sized land areas (approximately 283 000 and 269 000 km^2, respectively) that are divided by central mountain ranges separating eastern and western lowlands. Mainland Ecuador contains staggering biophysical gradients from the Pacific coastal plains, up through the Andean highlands (over 6200 m), and then down into upper reaches of the Amazonian basin. Strong environmental gradients also exist in Colorado, from the arid western canyons and mesas, up through the foothills and high peaks of the Central and Southern Rockies (over 4000 m), and down into the semi-arid grasslands to the east. Nonetheless, numerous geomorphic, biogeographic and hydroclimatic differences are apparent between tropical, coastally influenced Ecuador and temperate, mid-continental Colorado.

The ongoing development of new water infrastructure to secure municipal, hydroelectric and agricultural water supplies in Ecuador contrasts with the extensive history of logging, mining, ranching, irrigated farming and hydropower that have already transformed many rivers and their catchments in Colorado (Wohl, 2001). As in other emerging economies, the water resource regulations and protective national statutes that are taking shape in Ecuador require rapidly and transparently produced scientific guidance (Anderson *et al.,* 2011). The Instituto Nacional de Meteorologia e Hidrologia is working to expand the network of river gaging stations as well as the public availability of discharge records in Ecuador. However, both the number of stations and the duration of continuous, consistent measurement remain limited. For example, the monthly mean flows during some years between 1960 and 2010 are publically accessible for only 14 stations in Ecuador (http://www.serviciometeorologico.gob.ec/caudales-datos-historicos/). In contrast, the United States Geological Service offers real-time stream flow information for more than 300 sites in Colorado, and maintains historical statistics for more than 1200 gages, with several continuous records extending back to the early 20th century (http://waterdata.usgs.gov/co/nwis/current/).

Clustering approach

The analysis was conducted in the free and open-source R platform (3.2.2; R Core Team, 2015), and the script and data objects may be downloaded at http://figshare.com/s/4232e9bad11b11e4b70506ec4bbcf141. *Note this will link to a public DOI following review:* DOI 10.6084/m9.figshare.1309687).

We selected the recently released HydroBASINS data set to provide a spatially continuous set of polygons delineating the local contributing area for each reach in a drainage network based on a hydrologically corrected digital elevation model (Lehner and Grill, 2013). Whereas the precursor HydroSHED data sets were limited to large-scale basins; this new product offers a standardized set of consistently defined basins across several hierarchically nested spatial scales (using the Pfafstetter coding system). We opted to demonstrate the utility of this data set after confirming that the HydroBASIN delineations very closely resembled those generated independently by an alternative algorithm. The reach basin polygons encompassed the area upslope of the first channel junction in headwater catchments or the area draining to a junction-to-junction reach in subsequent downstream portions of the network. We conducted all analyses with the finest scale ‘Level 0’ (L0) reach basins but also examined the effect of using the coarser, larger ‘Level 8’ basins. Table I describes number and summary statistics for these units in both locations.

For each reach basin, we associated environmental attribute values derived from global, publicly available raster data sets recording elevation, climate, soils and land cover (Table II; Lehner *et al.* 2008; Hijmans *et al.,* 2005; Hengl *et al.,* 2014; Broxton *et al.,* 2014). We first performed basic terrain processing to yield topographic attributes and aggregated land cover from 17 to 6 types. We then examined pairwise scatterplots and correlations among attributes to select a set for clustering. In addition to elevation, slope, aspect and roughness, we included annual mean precipitation amount, the range of annual temperature, the month-to-month variation in precipitation, the percentages of clay, silt and coarse fragments, the estimated amount of organic soil carbon, soil pH and the percentages of forest, shrub, grassland, crop, urban and water/snow covers. Environmental feature values per reach basin were calculated as the mean value of the cells of each raster data layer within that basin polygon (i.e. as a zonal statistic for the reach basin).

We sought to identify the classification best supported by the data, while retaining the means to evaluate the relative strength of alternatives with fewer or greater numbers of classes. After exploring several algorithms including *k*-medoid and Gaussian mixture model routines, we chose to perform hierarchical agglomerative clustering (i.e. progressively grouping reach basins according to their multi-dimensional similarity). This well-studied procedure is mathematically straightforward, provides nested solutions as the number of clusters increases or decreases and accommodates skewed feature distributions. It has also been recommended and successfully implemented for inductive stream flow classifications based on time series data (Olden *et al.* 2012). Using Ward’s method (Ward, 1963), we aggregated reach basins via pairwise Euclidean distances calculated from scaled and centred features (i.e. after subtracting the feature mean and dividing by the standard deviation, we calculated the square root of the summed squared differences between observations).

As an alternative to more heuristic searches for points of inflection in the decline of group variation with more groups, the gap statistic provides a more consistent and theoretically grounded means to define the preferred number of groups given a range of cluster solutions (Tibshirani *et al.,* 2001; Maechler *et al.,* 2014; R function cluster::clusGap). The ‘gap’ refers to the difference calculated between the within-group variance resulting from a specific division of the data into k clusters and the variance for k clusters of a null distribution with the same number of observations and features. Taking the gap statistic value as a function of k, we adopted the rule proposed by Tibshirani *et al*. (2001) to select the smallest k producing a gap value greater than or equal to the value for $k+1$ minus one standard error (Maechler 2014). In effect, when gap values followed a saturating curve, this tended to select the class number at or near the earliest point of inflection. We also tested cluster stability by comparing the consistency of class membership across multiple bootstrap subsets of the segment basins, measured as the mean Jaccard similarity relative to the baseline cluster solution (Hennig, 2014; R function fpc::clusterboot). Greater cluster stability indicates that classifications are robust to reductions or substitutions in the sampled units (i.e. reach basins).

Table I. Reach basin counts and areas (km^2) for HydroBASINS polygon sets in Colorado and Ecuador at level 0 (fine) and level 8 (coarse).

Study region and resolution	*N*	Minimum	Median	Maximum
Colorado L0	2204	0.3	136	257
Colorado L8	426	0.3	535	5 467
Ecuador L0	2002	0.4	138	414
Ecuador L8	332	2.7	597	9 059

Table II. Environmental attribute data used in the analysis.

Data type	Features	Units	Resolution	Source
Topography	Elevation, aspect, slope, roughness	m	~90 m	HydroBASINS hydrologically conditioned DEM based on SRTM (http://www.hydrosheds.org)
Climate	Mean annual precipitation (Bio 12), CV monthly precipitation (Bio 15), Range annual temperature (Bio 7)	mm, C°	~1 km	BioClim: Hijmans *et al.* 2005 (www.worldclim.org)
Soils	Clay, silt and coarse fragments, estimated organic soil carbon, soil pH	%	~1 km	Soilgrids: Hengl *et al.* 2014 (www.soilgrids.org)
Land cover/use	Forest, shrub, grassland, crop, urban and water/snow; aggregated categories from MODIS types	%	0.5 km	Broxton *et al.* 2014 (http://landcover.usgs.gov/global_climatology.php)

In order to accommodate the possibility of highly resolved classifications, we allowed a maximum of 30 classes.

Comparison to existing classifications

McManamay *et al*. (2014) used a Gaussian mixture model algorithm to define 12 flow regime classes within the continental US according to measures of discharge magnitude, timing, frequency, duration and rate of change drawn from long-term streamflow records. As a preliminary assessment of the relationship between a classification based on environmental features and one based on observed flows, we selected those gages that overlapped with the classified area in Colorado for comparison. We eliminated gages with drainage areas larger than twice the maximum area of the L0 reach basins (514 km^2). The remaining 81 gages were located in a combination of 'reference' basins ($n = 46$; i.e. with little to no upstream hydrologic modifications such as dams, diversions, return flows or conversion of native vegetation), 'non-reference' basins ($n = 32$) and 'pre-dam' basins ($n = 3$). These gages were classified as intermittent flashy (IF), characterized by high intermittency and extended low-flow duration, stable groundwater (SSGW), with high baseflow and few fluctuations, and snowmelt (SNM1 and SNM2), with a distinct high-flow period, a winter minimum and relatively stable baseflow (SNM2 exhibited slightly higher baseflow and later high flow timing).

In addition, we compared the classification results in Ecuador with the freshwater ecoregions of the world (FEOW) that overlapped the national boundary (Abell *et al.* 2008; www.feow.org). In a notably comprehensive effort, several years of collaboration among more than 100 experts led to a classification based primarily on fish distribution information. The associated map may be the best widely available depiction of global ecohydrologic variation, and this research has supported subsequent assessments of environmental risk such as the threat to fish diversity posed by dams (Reidy Liermann *et al.*, 2012a).

RESULTS

Clustering performance

Gap statistic values increased rapidly with increasing numbers of class divisions before plateauing with smaller marginal increases (Figure 1). However, gap values continued to increase up to the maximum number of clusters considered in both Colorado and Ecuador. The finer-scale L0 reach basins supported more highly resolved classifications in both locations (indicated by vertical dashed lines in Figure 1), but both the selected number of clusters and the increase from L8 (coarse) to L0 (fine) solutions were larger in Ecuador (Figure 1B; preferred solutions occurred when the gap value increase from an additional group was within the standard error of a given number of clusters).

The overall stability of clustering solutions did not differ markedly between locations or with the scale of reach basins (Figure 2; stability was measured as the mean Jaccard similarity

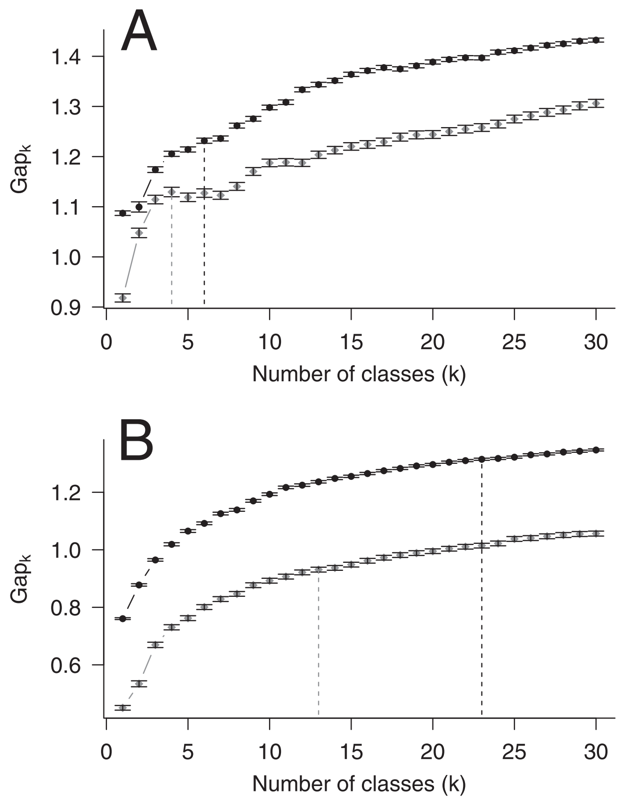

Figure 1. Gap statistic values with increasing number of classes for (A) Colorado and (B) Ecuador. Solutions for the finer-scale L0 reach basins (black circles) and coarser L8 (grey diamonds) are illustrated with the number of classes (dashed vertical lines) selected where the change in gap value from an additional group was less than or within the standard error of the value at a given number of clusters.

between the baseline cluster solution and 100 bootstrap replicates, with values approaching 1 for perfectly reproduced groupings). In Colorado, a class representing the drier, lower, eastern plains was most consistently distinguished across both sets of reach basin units. The wet, forested, eastern lowlands were similarly well defined in Ecuador for L0 and L8. However, the scale of basin units did influence the locations of lower cluster stability (lighter areas in Figure 2).

Colorado comparison

The best-supported cluster solution for the L0 reach basins in Colorado produced a system of six classes that reflected known hydro-climatic patterns (Figure 3A). These classes represented an area of arid canyonlands in the southwestern corner of the state (blue in Figure 3), mid-elevation shrublands to the east and west of the Continental Divide (red in Figure 3), slightly higher, wetter and more forested foothills (green in Figure 3), the central high-elevation mountainous areas (purple in Figure 3), the eastern Great Plains and San Luis Valley (orange in Figure 3) and peri-urban areas at the base of the Front Range corresponding to Colorado Springs, the Denver metro area and Ft. Collins (light blue in Figure 3).

The streamflow gages classified by McManamay *et al.* (2014) into flow regime types were located only within the mid-elevation shrubland, foothills and high-elevation mountainous classes (Table III). No flow types appeared to be located within plainly incorrect landscape classes. Both stable groundwater types (SSGW) occurred in the shrubland class, and the snowmelt types were limited to the foothills and mountains (with 56 of 66 SNM1 in the mountains). The five intermittent flashy (IF1) gages occurred in all three classes, but the two located in a mountain class were immediately adjacent to mid-elevation shrubland. The two least-common classes (southwestern canyonlands and Front Range peri-urban) were lost in the four class solution selected for the coarser L8 reach basins (not shown). As expected, the overall interspersion of classes was also reduced (e.g. compare Figure 2B to 2A), but the relationships with flow types were similar to those for the L0 classes.

Ecuador

The best-supported cluster solutions captured the major known gradients from the Pacific coastal lowlands, through the central Andean cordillera, and across to the Amazonian headwaters. However, despite reinforcing the importance of elevation differences, the classifications for both the L0 and L8 reach basins indicated the potential complexity of ecohydrologic processes in the Northern Andes with 23 and 13 classes, respectively (Figure 1). In particular, the classification of the L0 reach basins contrasted with the four FEOW zones within Ecuador (Figure 4), distinguishing boundaries at a much finer scale. A single eastern class based on environmental features included both the FEOW ‘Western Amazon piedmont’ and ‘Amazonas Lowlands’ zones, but more than a dozen L0 classes were within each of the North Andean Pacific Slope and Amazonas High Andes zones (and $n = 12$ and $n = 6$ L8 classes were present, respectively).

Catchment classes in Ecuador varied substantially in the environmental factors contributing to hydrological patterns (Figure 5, Table IV). In contrast to the increase in precipitation and percent forested cover with elevation in Colorado (Figure 3 lower panels), these key features showed evidence of greater divergence. For example, the neighbouring classes 4 and 5 exhibited similarly high forest cover, but the higher, steeper class 4 had less overall precipitation that was also more variable between months. Further to the south, classes 9 and 10 had comparable amounts of annual precipitation and clay soils, but the higher, steeper and more forested class 10 also displayed lower month-to-month variation in rainfall. Along the expansive coastal plain to the west of the Andes, less forest area, more clay and more cropped area distinguished class

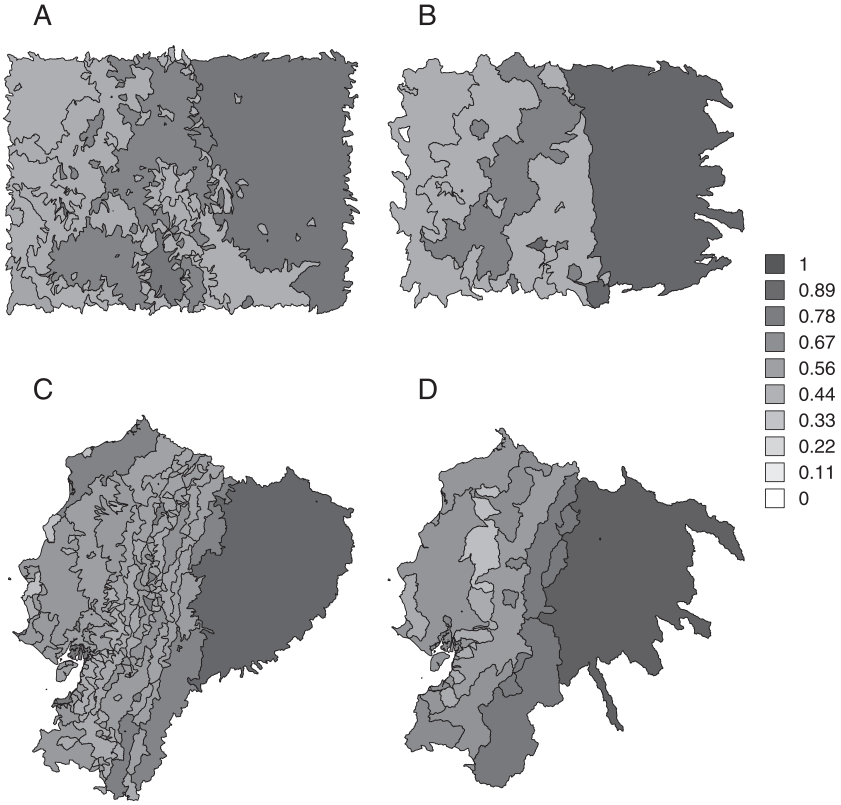

Figure 2. Stability across cluster solutions for Colorado (A, B) and Ecuador (C, D). Panels illustrate reach basin polygons dissolved according to class, with finer L0 (A, C) and coarser L8 (B, D) units. Darker colours indicate more stable classes (values approaching 1) that were more frequently regenerated across bootstrap replicates of clustering. The set of L8 reach basins extends over a slightly larger overall extent in both locations.

19 from the adjoining class 21, despite the relative topographic similarity in elevation and slope.

DISCUSSION

This research evaluated the number and spatial distribution of landscape classes having potentially distinct hydrologic conditions. The classes defined in Colorado reflect major hydroclimate divisions and accord with flow regime types based on observed discharge. Similarly, the classification results for Ecuador provide additional, finer-scale information to supplement the boundaries of existing global freshwater ecoregions. These findings suggest that an approach such as this one can usefully discern divisions for preliminary planning or regulatory processes in some data-scarce settings.

The study locations encompassed similar total areas, and the number and size of reach basins was comparable (Table I). However, the larger number of classes supported in Ecuador suggests that the method is sensitive to empirically greater spatial variation in environmental conditions in a tropical mountain setting relative to a temperate one. The range of values across reach basins was greater for 12 of the 18 features included in clustering, including precipitation amount and monthly variability, elevation, slope, roughness, soil properties and forest cover percent. In addition, 81 of 153 pairwise correlations among the features of L0 reach basins were lower in Ecuador (mean absolute values of Pearson r for all features, Ecuador = 0.29 versus Colorado = 0.36). However, as indicated by the steady increase in gap statistic values with greater numbers of divisions (Figure 1) and the moderate stability of most classes (Figure 2), the classifications in both locations provide plausible divisions over continuous functional gradients rather than serving to recover fundamentally distinct groups. This outcome highlights the importance of viewing classification results, from this or any other approach, as an evidence-based hypothesis rather than a definitive categorization of essential differences (Snelder and Booker, 2012).

Furthermore, if sharper transitions favour more clearly defined divisions, then locations with very gradual environmental gradients will be harder to classify well. Thus, the basic biophysical characteristics of some areas are likely to compound the challenge of classifying natural spatiotemporal variability, in addition to data accuracy and precision concerns. For example, very smoothly grading topography and precipitation amounts may hinder the subcategorization of lowland tropical forest or prairie steppe, as was evident in

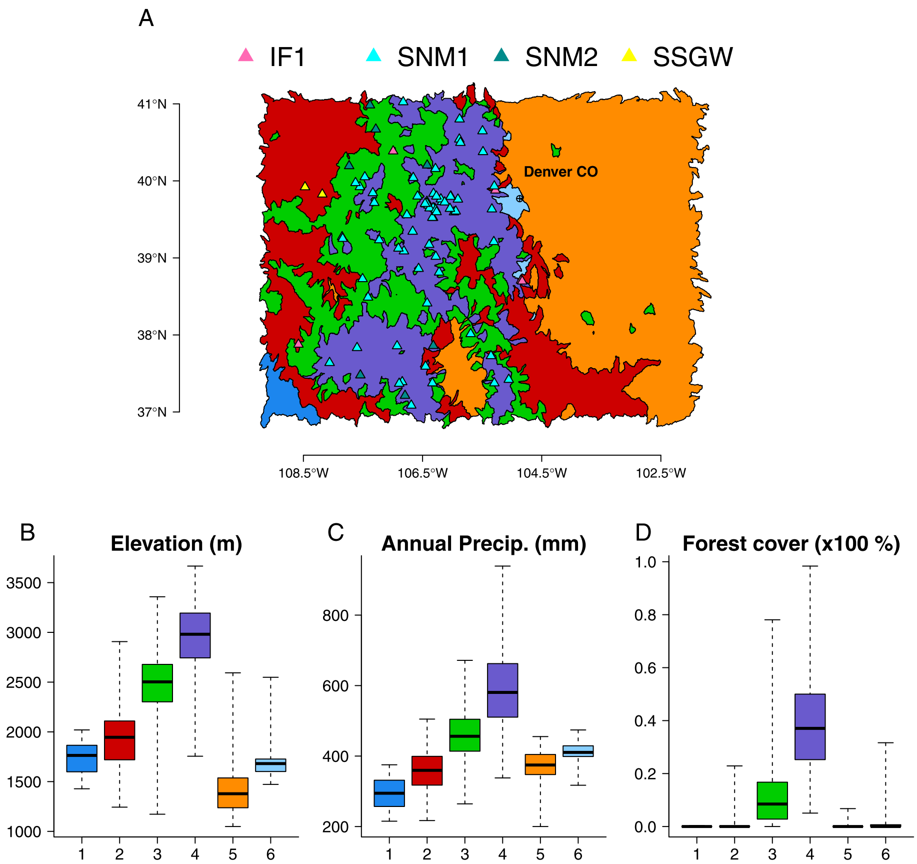

Figure 3. (A) Map of the best-supported classification using environmental features for L0 reach basins in Colorado. Triangles indicate the location of the long-term streamflow gaging stations used to develop the flow types used for comparison (IF = intermittent flashy, SNM1 and 2 = snowmelt, SSGW = stable groundwater). Colors of the 6 classes based on environmental features match those in panels B–D. (B–D) Per class distributions of values across reach basins for 3 of 18 environmental features used in clustering.

Table III. Correspondence between classes defined for Colorado according to environmental features and flow regime types based on streamflow records. Colours in parentheses refer to those in Figure 3. The included flow types were intermittent flashy (IF), snowmelt (SNM1 and SNM2), and stable groundwater (SSGW).

Landscape feature classes	IF1	SNM1	SNM2	SSGW
Mid-elevation shrublands (red)	2	0	0	2
Western Foothills (green)	1	10	5	0
Higher-elevation mountains (purple)	2	56	3	0

eastern Ecuador and Colorado. In this regard, the proposed nested classification approach has the advantage of spatially subdividing classes as classification levels are added. In contrast, non-hierarchical clustering methods may reconfigure the arrangement of classes in less intuitive ways (e.g. the transition from two to three classes may be less likely to involve a split within one of the two).

The specific goals of a classification effort will influence the most appropriate geographic scale of analysis, but this study suggests that an average reach basin area on the order of 100 km^2 (i.e. the scale of the HydroBASINS level 0 units) can support a well-resolved classification that remains tractable relative to intermediate political boundaries (e.g. states, districts or counties). Alternatively, an average area between 500 and 750 km^2 may remain suited to rapid analysis and yield a manageable number of classes over larger extents than considered here, while still capturing sufficiently detailed variation. Nonetheless, the results for the L0 and L8 reach basins indicated the need for further research to develop the relationship between the area of classified units and the best-supported number of classes. The 77% increase from L8 to L0 in Ecuador versus the 50% increase in Colorado confirmed the importance of controlling for consistent spatial units in any comparison of multiple regions. Future investigation could examine how this sensitivity to reach basin area differs across additional

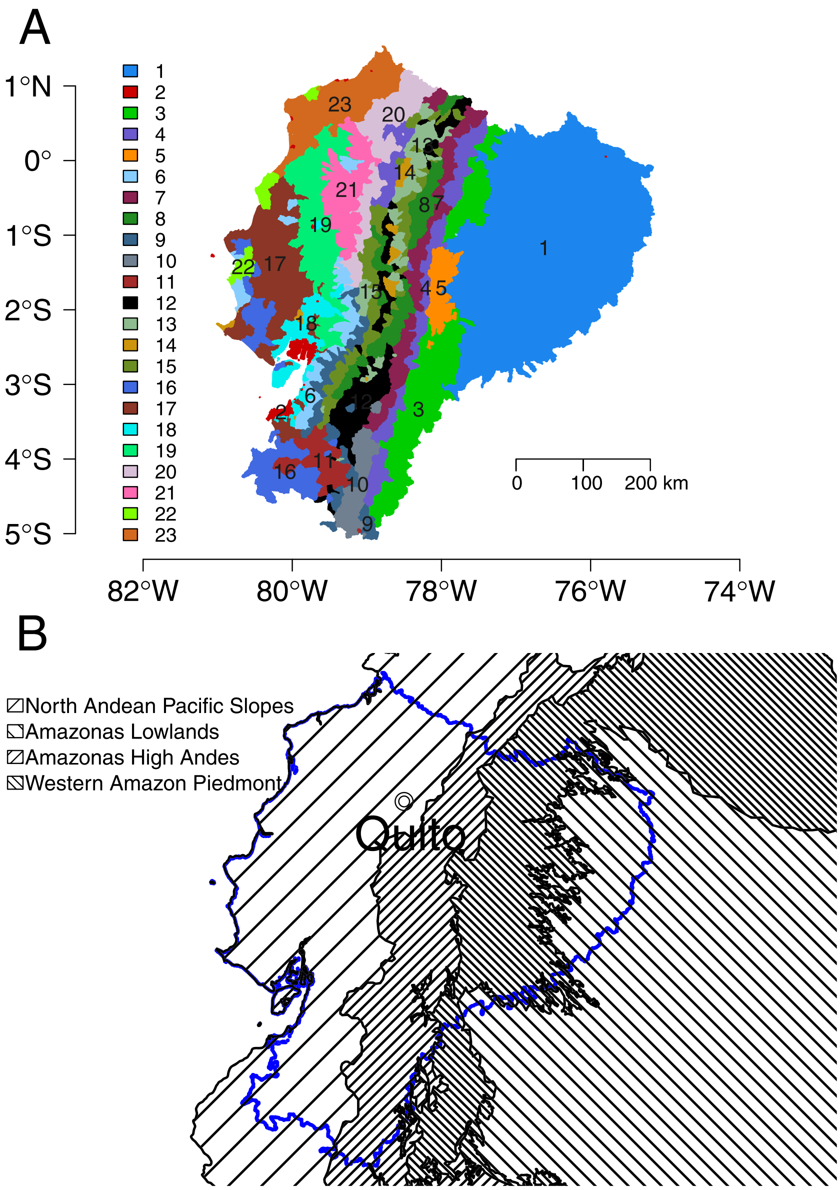

Figure 4. (A) Map of the best-supported classification using environmental features for L0 reach basins in Ecuador, and (B) the FEOW divisions present within the national extent. More than 10 classes overlapped with the North Andean Pacific Slopes for both the L0 and L8 sets of reach basins (see also Supplemental Figure 1).

regions, and whether it exhibits any consistent trends with latitude, ocean proximity or other major global features. The HydroBASINS data set greatly facilitates such research by making a considerable range of spatial units easily available, but another analysis could also examine the minimum flow accumulation threshold suitable for stream network definition and delineation of even finer reach basins.

The approach presented here is based on established statistical procedures and previous hierarchical environmental classifications shown to be effective and credible in data-rich areas (Snelder and Biggs, 2002; Wolock *et al.*, 2004; Snelder *et al.*, 2005; Sawicz *et al.*, 2011; McManamay *et al.*, 2012). However, we assume that the resulting classes best capture differences in riparian and in-channel conditions for reaches with upstream drainage areas within the reach basin unit (as distinct from the single 'mainstem' reach that forms the basis for the unit boundaries in non-first order reaches). Our decision not to calculate the values of environmental features over the entire upstream drainage area was motivated by interest in addressing smaller rivers and streams than have typically been the focus of global analyses. Defining sensible classes under the alternative approach of whole-watershed values also merits additional research into several key questions. For instance, what is the appropriate way to stratify spatial units when including cumulative upstream area? Are

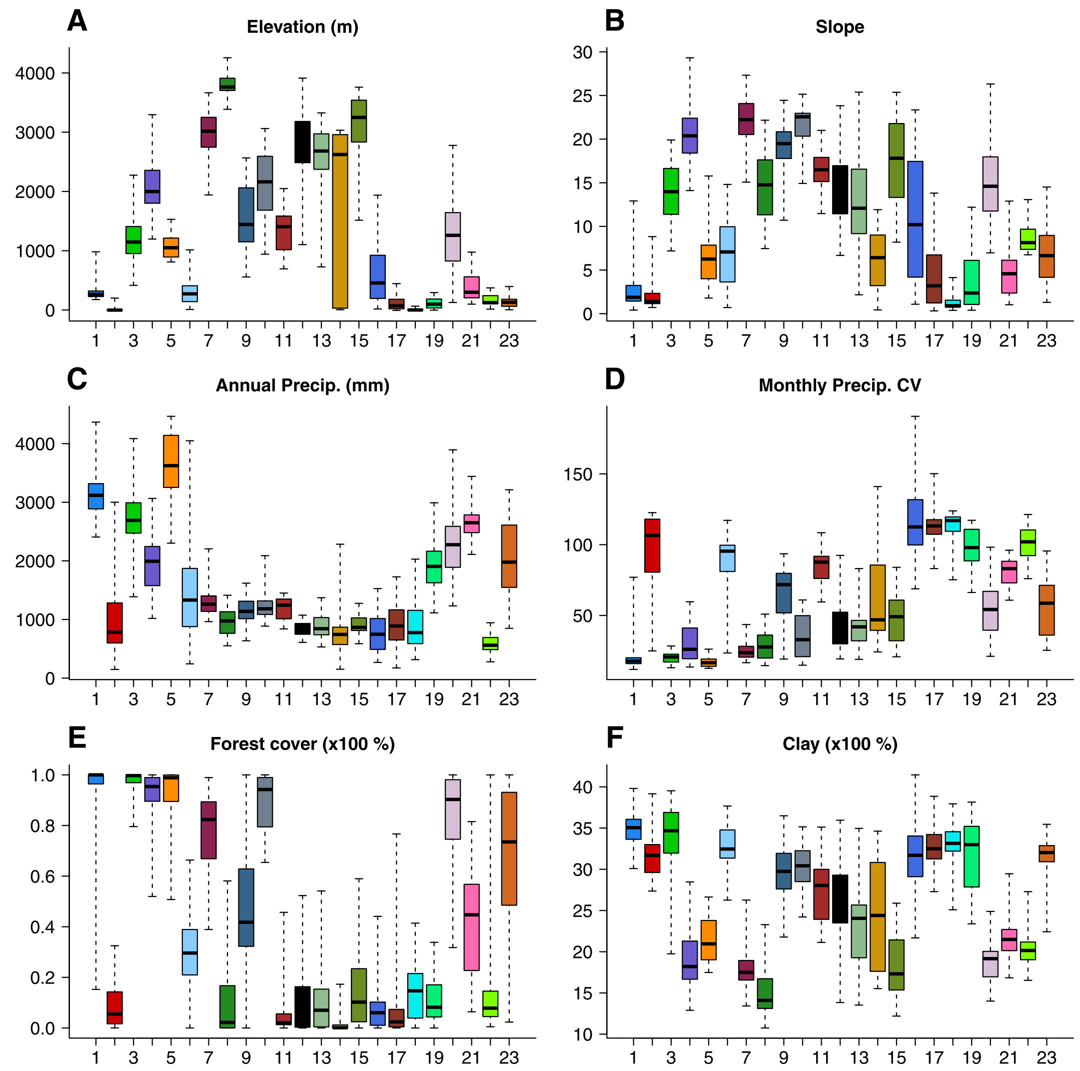

Figure 5. Distributions of reach basin characteristics across the 23 L0 classes for Ecuador. Colours and numbering follow Figure 4a.

distinct classifications justified for different Strahler orders, or should the drainage area itself be included as a clustering feature, or perhaps as the basis for a preliminary clustering serving as an alternative to traditional stream ordering? A spatially uniform weighting to compute the mean has the advantage of conceptual and technical simplicity, but a distance-weighted average may have greater relevance to many ecohydrologic variables of interest. If a distance-based method is preferable, should weights be based on network or Euclidean distance metrics or a combination that depends on the included variables? The emphasis on ‘local’ influences within the reach basin, rather than the cumulative biophysical attributes subject to river network structure, is an important caveat for interpreting this research. However, this method affords valuable conceptual and logistical simplicity while maintaining the scope to address larger river sections via the choice of HydroBASINS level.

Such considerations of network scale relate directly to management tasks involving environmental flow assessment and regulation. Environmental flow guidelines define limits for infrastructure permitting or operations that are intended to protect the flow regime attributes that favour native biota and ecosystem integrity. The development of a hydrologic classification has been proposed as an important step towards assessing environmental flows via a participatory process such as the Ecological Limits of Hydrologic Alteration (Poff *et al*., 2010). Such assessment is urgently needed amidst the intensifying pressure to develop water infrastructure in the Northern Andes (Finer and Jenkins, 2012). Indeed, this research was initiated after attempts to develop a hydrologic classification for Ecuador were stymied by the limited flow data available for direct analysis or to support rainfall-runoff modelling. The catchment classes produced from landscape features do not necessarily

Table IV. Catchment characteristics for selected features by class in Ecuador. Values indicate the median, minimum and maximum across the set of reach basins.

Class	Elevation (m)	Slope (%)	Precip. (mm)	Temp. range (C)	Clay (%)	Coarse Frag. (%)	Forest (%)	Grassland (%)	Cropped (%)
1	264 [176, 982]	1.9 [0.4, 12.9]	3118 [2407, 4367]	11 [11, 12]	35 [30, 40]	17 [3, 28]	100 [15, 100]	0 [0, 3]	0 [0, 81]
2	−2 [−15, 201]	1.4 [0.7, 8.8]	783 [148, 3001]	12 [8, 13]	32 [27, 39]	4 [1, 13]	6 [0, 32]	0 [0, 15]	0 [0, 24]
3	1147 [416, 2274]	14 [7.2, 19.9]	2690 [1388, 4087]	13 [11, 14]	35 [20, 40]	30 [16, 43]	100 [80, 100]	0 [0, 10]	0 [0, 18]
4	1998 [1197, 3296]	20.4 [14.1, 29.3]	1993 [1018, 3065]	13 [12, 15]	18 [13, 28]	33 [15, 56]	95 [52, 100]	1 [0, 30]	0 [0, 25]
5	1051 [811, 1530]	6.3 [1.8, 15.8]	3624 [2302, 4466]	11 [11, 12]	21 [17, 27]	19 [12, 31]	99 [51, 100]	0 [0,10]	0 [0,46]
6	272 [8, 1014]	7.1 [0.7, 14.8]	1331 [243, 4050]	12 [11, 12]	32 [26, 38]	9 [2, 19]	30 [0, 66]	11 [0, 26]	46 [7, 79]
7	3017 [1941, 3665]	22.2 [15.1, 27.3]	1261 [963, 2205]	12 [10, 13]	17 [13, 26]	52 [29, 72]	82 [39, 99]	11 [0, 43]	0 [0, 23]
8	3762 [3387, 4256]	14.8 [7.5, 22.2]	975 [551, 1413]	11 [10, 11]	14 [11, 23]	44 [31, 72]	2 [0, 58]	91 [38, 100]	1 [0, 24]
9	1443 [557, 2566]	19.5 [10.7, 24.4]	1139 [637, 1619]	12 [11, 14]	30 [22, 36]	20 [11, 36]	42 [0, 100]	15 [0, 58]	20 [0, 53]
10	2162 [941, 3062]	22.6 [14.9, 25.1]	1183 [886, 2089]	13 [11, 14]	30 [24, 35]	22 [14, 32]	94 [65, 100]	1 [0, 29]	1 [0, 31]
11	1406 [694, 2049]	16.5 [11.5, 21]	1244 [839, 1451]	14 [12, 14]	28 [21, 35]	14 [8, 18]	2 [0, 46]	58 [12, 96]	13 [0, 36]
12	2889 [1103, 3913]	14 [6.7, 23.8]	817 [609, 1074]	12 [11, 14]	27 [14, 36]	19 [7, 46]	5 [0, 52]	72 [19, 96]	10 [0, 43]
13	2683 [727, 3326]	12.1 [2.2, 25.4]	842 [531, 1371]	13 [12, 17]	24 [14, 35]	14 [4, 35]	7 [0, 54]	28 [0, 76]	41 [7, 100]
14	2623 [4, 3033]	6.4 [0.4, 11.9]	745 [151, 2284]	13 [8, 16]	24 [16, 35]	10 [1, 17]	0 [0, 17]	13 [0, 44]	13 [0, 50]
15	3249 [1515, 3760]	17.8 [8.2, 25.4]	865 [587, 1275]	12 [10, 13]	17 [12, 26]	30 [21, 45]	10 [0, 59]	62 [20, 96]	12 [1, 49]
16	456 [17, 1938]	10.2 [1.1, 23.3]	748 [267, 1528]	14 [11, 17]	32 [22, 41]	11 [6, 18]	6 [0, 44]	27 [4, 57]	10 [0, 55]
17	70 [−7, 445]	3.2 [0.3, 13.8]	888 [174, 1728]	12 [10, 14]	32 [27, 39]	7 [3, 15]	2 [0, 77]	29 [0, 90]	50 [7, 100]
18	0 [−13, 64]	0.9 [0.4, 4.1]	776 [314, 2030]	12 [11, 13]	33 [25, 38]	4 [2, 9]	15 [0, 41]	5 [0, 23]	24 [0, 56]
19	98 [−3, 294]	2.4 [0.4, 12.2]	1905 [1114, 2991]	11 [10, 12]	33 [23, 38]	3 [2, 8]	8 [0, 34]	5 [0, 26]	78 [48, 100]
20	1260 [127, 2777]	14.6 [7, 26.3]	2275 [1232, 3895]	11 [9, 13]	19 [14, 25]	15 [7, 31]	90 [32, 100]	1 [0, 34]	3 [0, 57]
21	299 [98, 977]	4.6 [1, 12.9]	2651 [2111, 3441]	11 [10, 12]	21 [17, 29]	3 [2, 10]	45 [6, 82]	6 [1, 28]	40 [14, 81]
22	125 [15, 373]	8.1 [6.7, 13.1]	561 [278, 944]	10 [8, 11]	20 [17, 27]	8 [7, 13]	8 [0, 100]	16 [0, 39]	45 [0, 74]
23	129 [4, 396]	6.6 [1.3, 14.5]	1978 [849, 3213]	9 [8, 10]	32 [22, 35]	10 [5, 14]	74 [2, 100]	1 [0, 12]	17 [0, 73]

quantify differences in specific flow regime features (e.g. the frequency, duration and timing of high-magnitude or low-magnitude discharge), but this research suggests that these classes do highlight the number and location of primary zones with similar hydrologic processes that give rise to these flow regime features. Thus, we envision that the classification system and maps resulting from this approach could clarify the need for more detailed environmental risk assessment or ground the design of additional monitoring networks. For instance, the classification in Ecuador could support planning to expand the system of streamflow gages. Under-represented classes in the current monitoring network might be targeted for additional flow monitoring (Figure 6).

If classes are assumed to correspond to distinct ecohydrologic regimes and thereby particular habitat conditions and biological communities – an assumption that requires further testing – then a classification such as this one could also help to guide planning for the expansion or revision of protected area networks for freshwater biodiversity (McManamay *et al.*, 2015). The capacity for

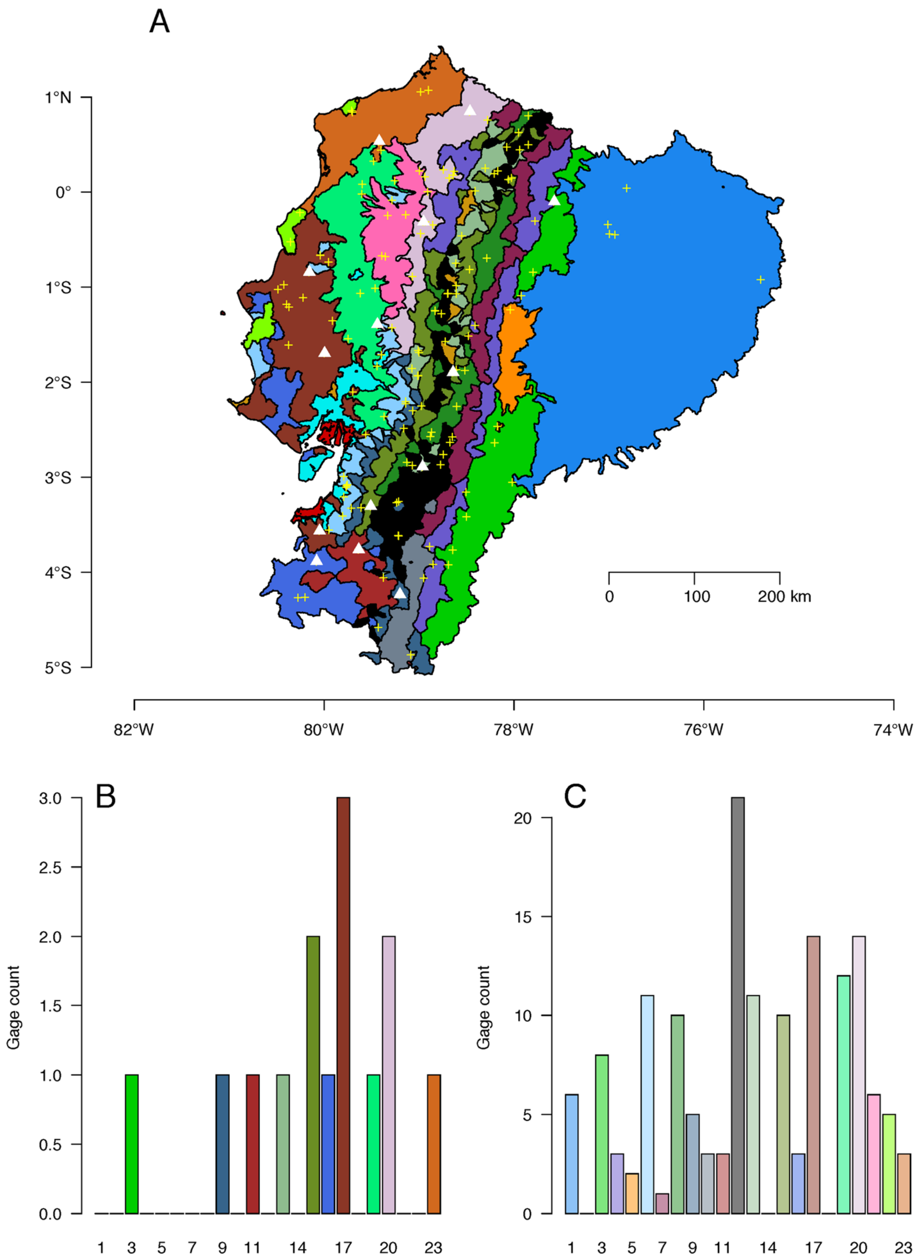

Figure 6. Location and number of Instituto Nacional de Meteorologia e Hidrologia streamflow gages relative to L0 classes in Ecuador The map in (A) illustrates locations with consistent, accessible data (white triangles; counts in panel B) and additional gaging stations (yellow crosses; counts in panel C) with unknown duration, consistency of measurement, and quality control.

rapid updating (e.g. with raster data drawn from climate or land cover scenarios of interest) could also permit an initial depiction of the hydrologic implications of anthropogenic impacts at a regional scale. Investigating changes in the number and distribution of classes under future scenarios might hold value as a way to visualize the loss, creation or relocation of niches in freshwater systems (Auerbach *et al.*, 2012). Further comparison with biogeographic data could also address questions of basic scientific interest, such as testing the hypothesis that greater diversity in ecohydrologic function predicts greater taxonomic or trait diversity in aquatic organisms.

This study also points towards opportunities to advance both deductive, geospatial catchment classification and the prediction of flows for ungaged locations in data-scarce areas. These research objectives are closely related because the environmental similarities that drive class groupings also support the transfer of rainfall runoff model parameter values or the statistical properties of streamflow. In particular, this protocol highlights the promise of automated data brokering and preprocessing to accelerate assessments. As new remote sensing instruments and distributed sensor networks are combined with the ongoing consolidation of data repositories, the improving quality and availability of public, global data sets are increasingly likely to support valid regional or national assessments in data-scarce areas. A number of recent projects have sought to develop end-user software for the estimation of 'reference' or 'natural' discharge properties via flow duration curves (Smakhtin and Eriyagama, 2008; Archfield *et al.*, 2013) and the comparison of multiple alternative prescribed flow regimes (Hickey *et al.*, 2014). These and other frameworks can benefit from dynamic connections to the kinds of environmental databases used in this study, permitting regularly updates to climate, soils or land cover data. For instance, Archfield *et al.* (2013) demonstrated a method to estimate discharge time series through statistical relationships between catchment properties and flow duration quantiles. Establishing automated services to provide a current representation of catchment properties could improve the hydrologic characterizations for various objectives.

The scripted workflow developed for this research reduces the time required to return a classification to a few hours, including download and processing time. However, very large areas with many reach basins are likely to require subdivision or additional parallelization, and regions with very irregular boundaries (e.g. archipelagoes and large inland water bodies) may require additional manipulation. Moreover, as noted earlier, our analysis was restricted to two mountainous regions in order to facilitate comparison. Thus, the approach should be applied with caution in other geographic settings (e.g. plains, deserts, temperate forests, etc.) where further testing and development is needed. Nonetheless, this work demonstrates the increasing viability of real-time analyses of ecohydrologic conditions in conjunction with planning or management activities.

CONCLUSION

Water resource managers and environmental planners are challenged to make informed, environmentally sound decisions that balance human and ecosystem needs. This challenge is especially acute where data-scarcity coincides with population growth, burgeoning water resource development and high freshwater species biodiversity. The classification framework outlined here affords a systematic and quantitative depiction of hydrologic variation at policy-relevant spatial scales. The method produced a classification system that was congruent with groupings based on flow records in the USA and was successfully applied to Ecuador. The proposed technique is flexible and can incorporate additional or alternative data sets, such as higher resolution, region-specific land cover or geochemical attributes. The nested solutions created by the hierarchical approach allow adjustment of the number of classes to suit a particular need without disrupting higher-order groupings. Although further testing in additional locations is needed, the resulting classifications could form a transparent, reproducible departure point for tasks such as the design of monitoring networks or the definition of environmental flow standards.

ACKNOWLEDGEMENTS

This research was stimulated by a workshop in Quito supported by the Atkinson Centre for a Sustainable Future at Cornell University and the Dimensions of Biodiversity EVOTRAC project (NSF 1045960, NSF 1046408, and NSF 1046991). We thank all workshop participants for their contributions to understanding Andean environmental flow needs. DA is grateful for support from The Nature Conservancy NatureNet fellowship. Heather Tallis and Spencer Meyer offered thoughtful suggestions that greatly improved the manuscript. We thank Jacob Bien and Erika Mudrak for valuable conversation regarding statistical options.

REFERENCES

Abell R, Thieme ML, Revenga C, Bryer M, Kottelat M, Bogutskaya N, Coad B, Mandrak N, Balderas SC, Bussing W, Stiassny MLJ, Skelton P, Allen GR, Unmack P, Naseka A, Ng R, Sindorf N, Robertson J, Armijo E, Higgins JV, Heibel TJ, Wikramanayake E, Olson D, Lopez HL, Reis RE, Lundberg JG, Perez MHS, Petry P. 2008. Freshwater ecoregions of the world: a new map of biogeographic units for freshwater biodiversity conservation. *Bioscience* **58**: 403–414. DOI: 10.1641/B580507.

Anderson EP, Encalada AC, Maldonado-Ocampo JA, McClain ME, Ortega H, Wilcox BP 2011. Environmental flows: a concept for addressing effects of river alterations and climate change in the Andes. In *Climate Change and Biodiversity in the Tropical Andes*, Herzog SK, Martínez R, Jørgensen PM, Tiessen H (eds). Inter-American Institute for Global Change Research (IAI) and Scientific Committee on Problems of the Environment (SCOPE): São José dos Campos, Brazil; 326–338.

Archfield SA, Kennen JG, Carlisle DM, Wolock DM. 2014. An objective and parsimonious approach for classifying natural flow regimes at a continental scale. *River Research and Applications* **30**: 1166–1183. DOI:10.1002/rra.2710.

Archfield SA, Steeves PA, Guthrie JD, Ries Iii KG. 2013. Towards a publicly available, map-based regional software tool to estimate unregulated daily streamflow at ungauged rivers. *Geoscientific Model Development* **6**: 101–115. DOI:10.5194/gmd-6-101-2013.

Archfield SA, Vogel RM. 2010. Map correlation method: selection of a reference streamgage to estimate daily streamflow at ungaged catchments. *Water Resources Research* **46**: W10513. DOI:10.1029/2009WR008481.

Auerbach DA, Poff NL, McShane RR, Merritt DM, Pyne MI, Wilding TK. 2012. Streams past and future: fluvial responses to rapid environmental change in the context of historical variation. In JA Wiens, GD Hayward, HD Safford, and CM Giffen (Eds.), *Historical Environmental Variation in Conservation and Natural Resource Management*. Wiley-Blackwell: Oxford; 232–245.

Broxton PD, Zeng X, Sulla-Menashe D, Troch PA. 2014. A global land cover climatology using MODIS data. *Journal of Applied Meteorology and Climatology* **53**: 1593–1605. DOI:10.1175/JAMC-D-13-0270.1.

Finer M, Jenkins CN. 2012. Proliferation of hydroelectric dams in the Andean Amazon and implications for Andes-Amazon connectivity. *PLoS ONE* **7**: e35126. DOI: 10.1371/journal.pone.0035126.

Hengl T, de Jesus JM, MacMillan RA, Batjes NH, Heuvelink GBM, Ribeiro E, Samuel-Rosa A, Kempen B, Leenaars JGB, Walsh MG, Gonzalez MR. 2014. SoilGrids1km — global soil information based on automated mapping. *PLoS ONE* **9**: e105992. DOI:10.1371/journal.pone.0105992.

Hennig C. 2014. fpc: flexible procedures for clustering. *R package version* **2**: 1–9.

Hickey JT, Newbold SJ, Warner AT. 2014. HEC-RPT — software for facilitating development of river management alternatives. *River Research and Applications: n/a-n/a* . DOI:10.1002/rra.2745.

Higgins JV, Bryer MT, Khoury ML, Fitzhugh TW. 2005. A freshwater classification approach for biodiversity conservation planning. *Conservation Biology* **19**: 432–445. DOI:10.1111/j.1523-1739.2005.00504.x.

Hijmans RJ, Cameron SE, Parra JL, Jones PG, Jarvis A. 2005. Very high resolution interpolated climate surfaces for global land areas. *International Journal of Climatology* **25**: 1965–1978. DOI:10.1002/joc.1276.

Lehner B, Verdin K, Jarvis A. 2008. New global hydrography derived from spaceborne elevation data. *Eos, Transactions, AGU* **89**: 93–94.

Lehner B, Grill G. 2013. Global river hydrography and network routing: baseline data and new approaches to study the world's large river systems. *Hydrological Processes* **27**: 2171–2186.

Maechler M, Rousseeuw, P, Struyf, A, Hubert, M, Hornik, K 2014. Cluster: cluster analysis basics and extensions. R package version 1.15.2.

McManamay R, Samu N, Kao S-C, Bevelhimer M, Hetrick S. 2015. A multi-scale spatial approach to address environmental effects of small hydropower development. *Environmental Management* **55**: 217–243. DOI:10.1007/s00267-014-0371-2.

McManamay RA, Bevelhimer MS, Kao SC. 2014. Updating the US hydrologic classification: an approach to clustering and stratifying ecohydrologic data. *Ecohydrology* **7**: 903–926. DOI: 10.1002/Eco.1410.

McManamay RA, Orth DJ, Dolloff CA, Frimpong EA. 2012. Regional frameworks applied to hydrology: can landscape-based frameworks capture the hydrologic variability? *River Research and Applications* **28**: 1325–1339. DOI:10.1002/Rra.1535.

Olden JD, Kennard MJ, Pusey BJ 2012. A framework for hydrologic classification with a review of methodologies and applications in ecohydrology. *Ecohydrology* **5**: 503–518. DOI: 10.1002/Eco.251.

Peñas F, Barquín J, Snelder T, Booker D, Álvarez C. 2014. The influence of methodological procedures on hydrological classification performance. *Hydrology and Earth System Sciences* **18**: 3393–3409.

Poff NL. 1997. Landscape filters and species traits: towards mechanistic understanding and prediction in stream ecology. *J N. Amer Benth Soc* **391–409**:

Poff NL, Richter BD, Arthington AH, Bunn SE, Naiman RJ, Kendy E, Acreman M, Apse C, Bledsoe BP, Freeman MC, Henriksen J, Jacobson RB, Kennen JG, Merritt DM, O'Keeffe JH, Olden JD, Rogers K, Tharme RE, Warner A. 2010. The ecological limits of hydrologic alteration (ELOHA): a new framework for developing regional environmental flow standards. *Freshwater Biology* **55**: 147–170. DOI:10.1111/j.1365-2427.2009.02204.x.

R Core Team 2015. R: a language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. http://www.R-project.org/.

Reidy Liermann CA, Nilsson C, Robertson J, Ng RY. 2012a. Implications of dam obstruction for global freshwater fish diversity. *BioScience* **62**: 539–548.

Reidy Liermann CA, Olden JD, Beechie TJ, Kennard MJ, Skidmore PB, Konrad CP, Imaki H. 2012b. Hydrogeomorphic classification of Washington State rivers to support emerging environmental flow management strategies. *River Research and Applications* **28**: 1340–1358. DOI:10.1002/rra.1541.

Richter BD, Davis MM, Apse C, Konrad C. 2011. A presumptive standard for environmental flow protection. *River Research and Applications* **28**: 1312–1321. DOI:10.1002/rra.1511.

Sawicz K, Wagener T, Sivapalan M, Troch PA, Carrillo G. 2011. Catchment classification: empirical analysis of hydrologic similarity based on catchment function in the eastern USA. *Hydrology and Earth System Sciences* **15**: 2895–2911. DOI:10.5194/hess-15-2895-2011.

Sawicz KA, Kelleher C, Wagener T, Troch P, Sivapalan M, Carrillo G. 2014. Characterizing hydrologic change through catchment classification. *Hydrology and Earth System Sciences* **18**: 273–285. DOI:10.5194/hess-18-273-2014.

Smakhtin VU, Eriyagama N. 2008. Developing a software package for global desktop assessment of environmental flows. *Environmental Modelling and Software* **23**: 1396–1406. DOI: 10.1016/j.envsoft.2008.04.002.

Snelder TH, Biggs BJF. 2002. Multiscale river environment classification for water resources management. *JAWRA Journal of the American Water Resources Association* **38**: 1225–1239. DOI:10.1111/j.1752-1688.2002.tb04344.x.

Snelder TH, Biggs BJF, Woods RA. 2005. Improved eco-hydrological classification of rivers. *River Research and Applications* **21**: 609–628. DOI:10.1002/rra.826.

Snelder TH, Booker DJ. 2012. Natural flow regime classifications are sensitive to definition procedures. *River Research and Applications* **29**: 822–838. DOI:10.1002/rra.2581.

Snelder TH, Lamouroux N, Leathwick JR, Pella H, Sauquet E, Shankar U. 2009. Predictive mapping of the natural flow regimes of France. *Journal of Hydrology* **373**: 57–67. DOI: 10.1016/j.jhydrol.2009.04.011.

Tibshirani R, Walther G, Hastie T 2001. Estimating the number of clusters in a data set via the gap statistic. *J Roy Stat Soc B* **63**: 411–423. DOI: 10.1111/1467-9868.00293.

Vorosmarty CJ, McIntyre PB, Gessner MO, Dudgeon D, Prusevich A, Green P, Glidden S, Bunn SE, Sullivan CA, Liermann CR, Davies PM. 2010. Global threats to human water security and river biodiversity. *Nature* **467**: 555–561. DOI:10.1038/nature09440.

Wagener T, Sivapalan M, Troch P, Woods R. 2007. Catchment classification and hydrologic similarity. *Geograph Compass* **1**: 901–931. DOI:10.1111/j.1749-8198.2007.00039.x.

Ward JH Jr. 1963. Hierarchical grouping to optimize an objective function. *Journal of the American Statistical Association* **58**: 236–244.

Wohl EE. 2001. *Virtual Rivers : Lessons from the Mountain Rivers of the Colorado Front Range*. Yale University Press: New Haven Conn.

Wolock DM, Winter TC, McMahon G. 2004. Delineation and evaluation of hydrologic-landscape regions in the United States using geographic information system tools and multivariate statistical analyses. *Environmental Management* **34**: S71–S88. DOI:10.1007/s00267-003-5077-9.

SUPPORTING INFORMATION

Additional supporting information may be found in the online version of this article at the publisher's web site.

