The CAMELS data set: catchment attributes and meteorology for large-sample studies

Nans Addor[1,*], Andrew J. Newman[1], Naoki Mizukami[1], Martyn P. Clark[1]

[1] Research Applications Laboratory, National Center for Atmospheric Research, Boulder, USA
* Now at the Climatic Research Unit, School of Environmental Sciences, University of East Anglia, Norwich, UK

Correspondence to: Nans Addor (n.addor@uea.ac.uk)

Abstract. We present a new data set of attributes for 671 catchments in the contiguous United States (CONUS) minimally impacted by human activities. This complements the daily time series of meteorological forcing and streamflow provided by Newman et al. (2015b). To produce this extension, we synthesized diverse and complementary data sets to describe six main classes of attributes at the catchment scale: topography, climate, streamflow, land cover, soil and geology. The spatial variations among basins over the CONUS are discussed and compared using a series of maps. The large number of catchments, combined with the diversity of the attributes we extracted, makes this new data set well suited for large-sample studies and comparative hydrology. In comparison to the similar MOPEX (Model Parameter Estimation Experiment) data set, this data sets relies on more recent data, covers a wider range of attributes and its catchment are more evenly distributed across the CONUS. This study also involves assessments of the limitations of the source data sets used to compute catchment attributes, as well as detailed descriptions of how the attributes were computed. The hydrometeorological time series provided by Newman et al. (2015b, https://dx.doi.org/10.5065/D6MW2F4D) together with the catchment attributes introduced in this paper (https://dx.doi.org/10.5065/D6G73C3Q) constitute the freely available CAMELS data set, which stands for Catchment Attributes and MEteorology for Large-sample Studies.

Short summary. We introduce a data set describing the landscape of 671 catchments in the contiguous USA: we synthesized various data sources to characterize the topography, climate, streamflow, land cover, soil and geology of each catchment. This extends the daily time series of meteorological forcing and discharge provided by an earlier study. The diversity of these catchments will help improving our understanding and modeling of how the interplay between catchment attributes shapes hydrological processes.

1 Introduction

Catchment attributes are descriptors of the landscape. Their interplay shapes catchment behavior by influencing how catchments store and transfer water. Because they are designed to capture and synthesize the multi-faceted composition of catchments, catchment attributes necessarily cover a wide range of features, such as the catchment climate, hydrology, land

cover, soil, geology, topography and river network. Over the last decades, catchment attributes have been developed in a variety of ways and are building the blocks of countless hydrological studies.

A fruitful research direction involves studies that explore interrelationships among catchment attributes. Key examples include
5 how the interaction of climate and topography influences vegetation productivity (Voepel et al., 2011), how aridity affects the angle of stream intersections, thereby constraining the shape of the river network (Seybold et al., 2017) or the extent to which land cover influences annual streamflow (Oudin et al., 2008) or evapotranspiration (Zhang et al., 2001). Catchment attributes are also a standard way to characterize catchment (dis)similarities and are consistently employed to develop catchment classifications (e.g., McDonnell and Woods, 2004; Wagener et al., 2007; Sawicz et al., 2011; Berghuijs et al., 2014).
10 Furthermore, there have been considerable efforts to use catchment attributes to reflect the structure of the landscape in models. One approach is to infer hydrological model parameter values from catchment attributes (Abdulla and Lettenmaier, 1997; Seibert, 1999; Hundecha et al., 2008; Samaniego et al., 2010; Hrachowitz et al., 2013), with the parallel objectives of accounting for landscape characteristics in an explicit way (not only implicitly by calibration), and of implementing hydrological models in ungauged basins. Another approach is to base not only parameter values, but also the choice of model
15 structure on catchment attributes (Clark et al., 2011; McMillan et al., 2011; Fenicia et al., 2014). Both approaches provide guidance on how to deal with geophysical characteristics that vary dramatically within the model domain, for instance in the context of continental scale modeling.

Although catchment attributes are routinely used when working with a handful of catchments, there is a growing recognition
20 that a large sample of catchments can provide insights that cannot be gained from a small sample (Gupta et al., 2014). Large-sample data sets enable us to concentrate on catchment similarities, and on the formulation of conclusions that are valid for a large number of (gauged and ungauged) catchments. Individual catchments can then be considered to be part of a continuum of catchment attributes, which vary in space along several gradients (such as aridity or soil depth). Working with a large number of catchments enables us to progressively move along different gradients, and to better disentangle the effects of
25 catchment attributes on catchment behavior. This is particularly useful for comparative hydrology, i.e., to identify how similarities and differences between locations influence eco-hydrological processes (Falkenmark and Chapman, 1989; Troch et al., 2009; Thompson et al., 2011; Harman and Troch, 2014). Further, large-sample hydrology opens new opportunities for data analysis, and for instance makes it possible to explore interrelationships between catchment attributes on the basis of their spatial patterns, as exemplified later in this study using map comparisons.
30
Several data sets of catchment attributes for large-sample hydrology now exist (see the review by Gupta et al., 2014). The large-sample data set introduced in this paper is an extension of the Newman et al. (2015b) data set, referred to as N15 in continuation. N15 covers 671 catchment in the contiguous USA (CONUS), for which it provides daily meteorological forcing from three data sets, Daymet (Thornton et al., 2012), Maurer (Maurer et al., 2002) and NLDAS (Xia et al., 2012), as well as

daily streamflow measurements from the United States Geological Survey (USGS). All those catchments have 20 years of continuous discharge record from 1990 to 2009 and are minimally impacted by human activities (see Section 2.1 in Newman et al., 2015). Here we cover the same catchments and provide additional quantitative estimates of a wide range of catchment attributes. We named this extended N15 data set the CAMELS data set, which stands for Catchment Attributes and

5   MEteorology for Large-sample Studies.

Section 2 explains the motivations to extend the N15 data set. Sections 3 to 8 present six classes of basin characteristics: topographic characteristics, climate indices, hydrological signatures, and land cover, soil and geology characteristics, respectively. These six sections are organized using the following structure. We first provide some research background on

10   this class of basin characteristics, introduce the attributes we selected and explain the reasons behind their selection. Since these attributes are well established, we briefly introduce them in the main text and provide further details in tables, which contain units and abbreviations for the attributes, as well as references to the equations and data sets used for their computation. We follow by discussing the spatial variations of these attributes across the CONUS and by assessing their main limitations of the source data sets. Section 9 compares the CAMELS data set to the MOPEX data set (Model Parameter Estimation

15   Experiment, Duan et al., 2006; Schaake et al., 2006), another large sample of catchments for the CONUS. Section 10 discusses the online availability of the CAMELS data set and possible future extensions. Conclusions are presented in Section 11.

2   Motivations to extend the Newman et al. (2015b) data set

In creating the CAMELS data set we seek to achieve the following objectives:

1. Make a wide range of geophysical data sets available and comparable at the catchment scale: We compiled
20   complementary catchment attributes from diverse data sources and synthesized them into a single coherent data set.
           These attributes have been available separately for some time, but comprehensive multivariate catchment scale
           assessments have so far been difficult, typically because disparate data sets have different spatial configurations, are
           stored in different archives, and use different data formats. By creating catchment scale estimates of these attributes,
           we simplify the assessment of their interrelationships.

25   2. Summarize meteorological forcing and discharge daily time series: We derived climate indices and hydrological
           signatures using the daily time series from N15. We selected climate indices and hydrological signatures that reduce
           the dimensionality of the hydroclimatic data sets, while preserving most of their information content. In other words,
           daily time series are rich in information, but summarizing this information makes catchment comparison easier.

3. Characterize catchment land cover, soil and geology: We leveraged data sets not used in N15 in order to describe the
30   land cover, soil and geology of each catchment. The attributes we extracted are commonly used to explore catchment
           behavior and to support parameter estimation for hydrological and land surface models. A goal is to better assess how

well those data sets capture the landscape features that matter for the storage and transfer of water across the landscape.

4. Define limitations in catchment attributes: Our intention is not only to provide quantitative estimates of diverse catchment characteristics, but also to explore and discuss limitations of those estimates. Catchment attributes are uncertain for different reasons, so we are providing metadata of different kinds (e.g., the difference in basin area estimated using different data sources, or the fraction of soil poorly characterized). Our aims are i) to contribute to raise awareness on uncertainties in geophysical attributes, which are frequently considered in a purely deterministic way, and ii) to facilitate catchment selection based on the reliability of their attributes.

Ensure spatial consistency across the domain: We reduce the risk of generating artificial regional variations by using only data sets covering the entire CONUS, and not different data sets for different parts of the domain.

For most variables and catchments, the spatial resolution of the source data set (e.g., the remote-sensed land cover characteristics) is smaller than the catchments, making upscaling necessary. By default, the upscaling was done using the arithmetic mean, except where indicated otherwise.

3   Location and topography

Location information and topographic indices were extracted for each catchment by N15 (Table 1). We display these attributes on maps to introduce the main topographic features of the CONUS. Elevation obviously exerts a key control on catchment behavior (Figure 1a), as it strongly influences a wide range of catchment attributes that we present in this paper, such as soil depth, land cover, the fraction of the precipitation falling as snow or streamflow seasonality. Figure 1b illustrates that the eastern half of the CONUS is, with the exception of the Appalachian Mountains, much flatter than its western counterpart. Figure 1c shows the spatial distribution of catchment size and highlights that there are some large catchments: five catchments have an area greater than $10,000 km^2$, and four of those are located in the Great Plains. Since we compute the catchment average of every attribute presented in this paper, it is important to keep mind that those catchment averages become less meaningful as the catchment area increases. In the context of hydrological modeling, the larger the catchment, the greater the need to account for spatial heterogeneity using some kind of spatially distributed representation.

As explained earlier, our aim is to reveal weaknesses in catchment characteristics and to discuss the impacts of such weaknesses for hydrological modeling. One way to do so is to compare different estimates of the same quantity, for instance catchment area. Two methods were used to determine the contours of each catchment, Geospatial Fabric (Viger, 2014; Viger and Bock, 2014) and GAGES II (Falcone, 2011). The polygons from Geospatial Fabric were instrumental to produce the N15 data set, since they were used to clip the gridded forcing data sets and the digital elevation model (from which elevation bands were derived), and importantly, they were used to estimate the area of each catchment, which enabled the conversion of discharge at the catchment outlet to average runoff depth over the catchment. It is hence essential to determine if the area computed from

the Geospatial Fabric polygon is reliable. We compared it to the area computed using the GAGES II data set and computed the absolute relative error between the two estimates. In eight catchments, the error is greater than 100% (red dots in Figure 1d), and in 62 catchments the relative error is greater than 10% (red and orange dots). Several of these catchments are located in the Great Basin and in California where the Geospatial Fabric had difficulty identifying watershed boundaries. Additionally,

5   the Geospatial Fabric was not designed to exactly replicate basin area above gauging locations, rather its development focused on continental scale hydrologic modeling, thus some basin area discrepancies are inherent in the development of the Geospatial Fabric. We recommend not using catchments with large error discrepancies with GAGES-II, as they are most likely erroneous in the Geospatial Fabric (e.g., Bock et al., 2016). Note that in general, catchment delineation is more challenging in flat areas, but here errors in flat areas are relatively well contained, except in Florida.

10   **4   Climate indices**

4.1   Data and methods

Climate indices were derived using meteorological forcing data from N15. N15 includes forcing from three data sets (NLDAS, Maurer and Daymet) but for the computation of the indices, only Daymet data were used. All the climate indices and hydrological signatures (Section 5) were computed for the period 1989/10/01-2009/09/30 (hydrological years 1990 to 2009).

15   The choice of this period was based on the proportion of missing daily discharge measurements (the forcing time series were all extracted from gridded data sets and are all complete). We consider this period to be long enough to derive climatological indices (in particular when rare events are characterized) and short enough to be little impacted by the lack of daily discharge measurements at the beginning and end of the period covered in N15 (1980-2014, see Figure 2).

20   There is a wide range of climatic indices in the literature. We selected indices with the goal to synthetize this myriad of possibilities and to provide direct support to the study of hydrological processes (Table 2). These indices characterize dry periods, high precipitation events and the baseline over two time scales: the daily time scale (e.g. frequency of high precipitation events) and the seasonal time scale (e.g. the proportion of precipitation falling as snow).

25   At the seasonal time scale, we computed three indices: aridity, the fraction of the precipitation falling as snow, and the seasonality and timing of precipitation. These three indices were previously used for the classification of 321 catchments across the CONUS and were shown to provide relevant insights into the relationship between catchment behavior and their physiographic characteristics (Berghuijs et al., 2014, note that we use slightly different formulations of these indices, see Table 2). Aridity is defined as the ratio of mean annual potential evapotranspiration over the mean annual precipitation. The

30   occurrence of snow was estimated for daily time steps using a temperature threshold of 0°C. The seasonality and timing of precipitation are combined into a single metric, which relies on sine curves representing the annual cycle of precipitation and temperature. Note that sine curves do not necessarily provide a good fit to the annual precipitation cycle, for instance in areas

experiencing a strong annual cycle and multiple consecutive months with low precipitation, such as California (see Berghuijs and Woods, 2015 for a solution to this issue), yet they enable a first-order characterization of the dominant climatological features of diverse locations, which is useful for studies such as this one. These three seasonal indices provide a good overview of the mean and seasonal climatic conditions, but do not explicitly consider dry periods and intense precipitation events, which

5   occur at different time scales and are key drivers of droughts and floods. To fill this gap, we considered the frequency of dry days and high precipitation events, as well as the mean duration of these events and determined the season during which most of the high precipitation events and dry days occur. This provides some insights into the precipitation regime (convective or stratiform) and phase (liquid or snow).

4.2   Results and discussion

10   The annual precipitation cycle is strongest over the Pacific Coast (maximum in winter), over the northern Great Plains and Florida (maximum in summer) and weakest along the Atlantic Coast (Figure 3a). The fraction of precipitation falling as snow is highest over the Rocky, Cascade and Sierra Nevada mountain ranges, followed by the Northeast and the Great Lakes Region (Figure 3b). Aridity is the highest over the Southwest, High Plains and Great Plains, when in contrast, the Northwest, Northeast and the Appalachians are the most humid regions (Figure 3c). High precipitation events occur most frequently in winter along

15   the Pacific Coast (Figure 3f) and are relatively long-lasting (Figure 3e), which reflects their large (synoptic) scale nature. In contrast, summertime convective systems (e.g., mesoscale systems) over the High Plains, Great Plains and the Upper and Middle Mississippi Valley generate the most frequent high precipitation events. In the band stretching from Louisiana to Georgia, high precipitation events are most frequent in winter, as the result of the intense extratropical cyclone activity. The frequency of dry days (Figure 3g) is closely related to aridity (Figure 3c). Catchments located in the region stretching from

20   California to Texas typically experience the longest periods of successive dry days, while those in the Northeast are at the other end of the spectrum (Figure 3h). Dry days are particularly frequent in summer west of the Rocky Mountains, in winter in the Great Plains and Mississippi Valley and in autumn in the Atlantic Coast States (Figure 3i).

5   Hydrological signatures

5.1   Data and methods

25   Hydrological signatures were chosen using a similar rationale as for climate indices: we aimed to capture the hydrological baseline, as well as low flow and high flow events. All signatures were computed using daily discharge time series retrieved by N15 from the USGS for the period 1989/10/01-2009/09/30 (Figure 2).

We selected signatures from the set that Sawicz et al. (2011) used to explore the similarity between 280 catchments in the

30   Eastern USA and classify them (Table 3). The runoff ratio indicates how much of the long-term precipitation leaves the catchment as streamflow, thereby reflecting losses to evapotranspiration and groundwater. We use the slope of the flow

duration curve to characterize streamflow variability: steeper flow duration curves define greater variability over the year. The slope is computed between the log-transformed 33rd and 66th streamflow percentiles. In intermittent streams, the frequency of days with no flow can be greater than 33%, so that Q33 = 0 mm/day. Since the logarithm cannot be extracted, the slope of the flow duration curve is undefined. The contribution of baseflow to the total discharge is estimated by the baseflow index

5 computed by hydrograph separation using a digital filter implemented by Ladson et al. (2013). It has to be recognized that the technique used for the separation influences the estimated baseflow index (see Beck et al., 2013; Ladson et al., 2013 for recent examples), yet hydrograph separation can provide valuable insights into catchment behavior (e.g., Harman et al., 2011) and the baseflow index has proven to be a useful variable to compare and classify large samples of catchments (e.g., Sawicz et al., 2011; Beck et al., 2016). Further, catchment response to a change in precipitation, which is particulary relevant in the context

10 of climate change (e.g., Vano et al., 2015), was evaluated by computing the elasticity between annual precipitation and discharge. Finally, we characterized discharge seasonality using the date of the center of mass. This indicator is frequently used to quantify the impacts of climate change on the hydrology of snow-dominated catchments (e.g., Court, 1962; Stewart et al., 2005; Addor et al., 2014). The center of mass dates have been shown to occur earlier, as temperature increases can force both an earlier onset of snow melt and a higher proportion of precipitation falling as rain.

Since the hydrological signatures introduced so far do not explicitly consider low and high flow events, we defined high and low flow days using thresholds based on the median and mean daily flow, respectively (Clausen and Biggs, 2000; Olden and Poff, 2003; Westerberg and McMillan, 2015). We computed the average duration and average frequency of high and low flow events. We also extracted 5th and 95th percentiles (Q5 and Q95, respectively) from the flow duration curve to characterize those

20 events.

5.2 Results and discussion

The mean annual discharge and runoff ratio are strongly correlated (Figures 4a and b) and present clear similarities to catchment aridity (Figure 3c). In the Great Plains, where the evaporative demands exceed available precipitation (aridity > 1), more than 80% of the precipitation is evaporated (runoff ratio < 0.2) and the mean annual discharge is often as low as 0.3

25 mm/day. In contrast, in the Pacific Northwest, precipitation is often twice as high as PET (aridity < 0.5). Both the runoff ratio and the mean annual discharge are higher in the Pacific Northwest than in the Northeast, as can be expected from the seasonality of precipitation, which peaks in winter in the Pacific Northwest (Figure 3a). Most of the discharge flows during the first half of the year in the Pacific Northwest (mean date of center of mass < 156, Figure 4c). Streamflow is in contrast delayed by snow accumulation in the Rocky Mountains, it is also late in the Midwest (in part because of the seasonality of the

30 precipitation, which peaks in summer), and in contrast early in the band stretching from eastern Texas to South Carolina (which is consistent again with the seasonality of precipitation). Similarities exists between the patterns of the slope flow of the duration curve and the baseflow index (Figure 4d and e), with lower baseflow index and higher slopes both indicating flashy

catchments, a clear example being the area stretching from east Kansas to Kentucky. Finally, at the annual scale, the discharge of more arid catchments tends to react more strongly to changes in precipitation (Figure 4f, see also Harman et al., 2011).

The frequency of low and high flow events is correlated (Figure 4g and j) and by definition, both frequencies are low in
5 catchments with a low slope of the flow duration curve (Figure 4d). High flows are least frequent and the most short-lived in the Pacific Northwest and in the Appalachian Mountains, and when they occur their absolute discharge is higher than in other regions (Figure 4i). Q5 is more than ten times higher in the Pacific Northwest and in the Appalachian Mountains than in the most arid catchments, which reflects the capacity of these humid catchments to sustain baseflow.

10 Note that even though spatial patterns emerge from the maps in Figure 4, they tend to be less smooth than those of climate indices shown in Figure 3. In other words, there can be some strong variations over short distances, for instance in the slope of the duration curve or in signatures related to extreme (high and low) streamflow conditions. Plausible explanations are that i) hydrological signatures are the end result of the interactions between several non-linear processes (as opposed to the smaller number of processes controlling for instance the fraction of precipitation falling as snow), ii) hydrological signatures are
15 sensitive to uncertainties in discharge measurements (Westerberg et al., 2016), which we suspect to contribute to sudden variations over short distances in the maps of particularly sensitive signatures, such as the slope of duration curve or signatures related to extreme streamflow conditions.

6   Land cover

6.1   Data and methods

20 We considered two key indicators of vegetation density: the leaf area index (LAI) and the green vegetation fraction (GVF), which approximates the vertical and horizontal density of vegetation, respectively. We used the 1km land cover products derived from the Moderate Resolution Imaging Spectroradiometer (MODIS) data to estimate their climatological monthly values over 2002-2014. LAI is defined as the one-sided green leaf area per unit ground area in broadleaf canopies and as half the total needle surface area per unit ground area in coniferous canopies. We extracted the maximum monthly LAI, which can
25 be used to constrain the maximum evaporative capacity and vegetation interception capacity in models. Seasonal variations in LAI are principally related to trees growing and shedding their leaves. To quantify these variations, we computed the difference between the maximum and minimum monthly LAI. In absolute terms, these variations are highest in areas of deciduous broadleaf forest. The GVF can be used in models to estimate the proportion of each grid cell covered by vegetation (1 minus the GVF gives the fraction of the grid cell from which evaporation occurs directly from the soil). Variations in the GVF are
30 particularly high for croplands, as the result of the growing and harvesting of the crops. Like for the LAI, we extracted the maximum monthly value of the GVF and the difference between the maximum and minimum monthly value.

Additionally, we included the land cover class based on the International Geosphere-Biosphere Programme (IGBP) classification (Belward, 1996) derived from MODIS data. For each catchment, we defined the dominant land cover class as the most frequent class based on all the grid points fully or partially contained in the basin (each grid cell was weighted based on how much of it was contained within the basin boundaries). The fraction of the catchment covered by the dominant class

5   is an indicator of the representativeness of the dominant class for the whole catchment.

Finally, based on the IGBP land cover class of each grid point, we approximated the root-depth distribution based on Zeng (2001). The distribution is estimated using a two-parameter equation, the value of these parameters being dependent on the IGBP land cover. The root fraction decreases exponentially with soil depth: the depth of the soil layer encompassing the top

10   50% of the root system is typically between 0.12 and 0.26m depending on the land cover, and for the top 99% of the root system, this depth is typically between 1.4 and 2.4m and is often named "rooting depth". We computed the root-depth distribution for each grid point based on its land cover. We then extracted the values associated with the following percentiles: 10, 25, 50, 75 and 99%. For each percentile, the catchment average was estimated using the arithmetic mean. Table 4 provides the complete list of land cover attributes that we considered.

15   **6.2   Results and discussion**

The maximum LAI and GVF are highly correlated (Figure 5a and d, see also the mean value for each land cover class in Figure 5k), which reflects that short vegetation tends to be sparse and forests of taller trees tend to be dense, but could also indicate that the MODIS data used to compute these two fields do not enable us to fully differentiate between vertical and horizontal vegetation density. These two fields are similar to that of the fraction of forest (Figure 5c, positive correlation) and aridity

20   (Figure 5c, negative correlation), with arid catchments typically associated with a lower LAI and lower GVF. Note that because the catchments selected are minimally impacted by human activities, none of them are classified as urban (Figure 5j).

The amplitude of the seasonal variations of LAI are strongly linked to LAI max (Figure 5a and b). Overall, catchments dominated by land cover classes with a high LAI (e.g., deciduous broadleaf forest or mixed forest, Figure 5j) tend to experience

25   a significant increase and drop of LAI, a clear exception to this rule being evergreen broadleaf forests (Figure 5k). The seasonal variations of GVF are particularly high for croplands, which is expected and reflects the growing-harvesting cycle. However, note that there are also important seasonal changes in the catchment dominated by the deciduous broadleaf forests (Figure 5k), although the horizontal tree density does not change significantly. Again, this suggests that the MODIS data does not enable to fully differentiate between vertical and horizontal vegetation. Users might hence decide to consider LAI only to summarize

30   seasonal land cover variations.

To explore spatial variations in rooting depth, we used aridity and depth to bedrock (introduced in Section 7.1). Figure 5l shows that in water-limited catchments (aridity > 1) the rooting depth increases with aridity, which can be interpreted as a sign

that trees increase their root zone storage capacity to compensate for the overall lack of water. In more humid catchments (aridity < 1), shallow soils seem to constrain the vertical development of roots of tall trees like evergreen needleaf forests and deciduous broadleaf, and in contrast, mixed forest and evergreen broadleaf forest can develop deeper roots. These hypotheses illustrate how the CAMELS data set enables the joint exploration of diverse attributes for a large number of catchments.

7   Soil characteristics

7.1   Data and methods

The soil characteristics we derived are principally based on the State Soil Geographic Database (STATSGO) data set post-processed by Miller and White (1998). Miller and White (1998) discretized the top 2.5m of soil into 11 layers, whose thickness increases with depth (from 5cm for the two top layers to 50cm for the three deepest ones). For each layer, they relied on the original STATSGO data to determine the dominant soil texture class. They considered a total of 16 classes, the 12 standard United States Department of Agriculture (USDA) soil texture classes plus four additional nonsoil classes characterized as organic material, water, bedrock and other.

We estimated the saturated hydraulic conductivity and porosity (saturated volumetric water content) of each layer using the multiple regressions relying on sand and clay fraction originally proposed by Cosby et al. (1984) and now commonly for land surface modeling (e.g., Lawrence and Slater, 2008). For organic material, we used default values for the saturated hydraulic conductivity and porosity based on Lawrence and Slater (2008). Then for each SATSGO polygon, we computed the average of each soil characteristic (see list in Table 5) over the top 1.5m of soil using the following weighted mean:

$$X_p = \sum_{i=1}^{i=9} X_i\, T_i \Big/ S_{depth}$$

where X_p designates the mean value of the variable X over the 1.5m of soil (nine top layers, see points 1 and 2 below), X_i is its value over layer i, T_i is the thickness of layer i and S_{depth} is the cumulated depth of the layers. Then for each catchment, we computed the weighted mean of the soil characteristics of the STATSGO polygons within the catchment, the weight being the fraction of the catchment covered by each polygon. For hydraulic conductivity, the harmonic mean was used instead of the arithmetic mean for the averaging along each soil column and across the catchment (see Samaniego et al. (2010) for a discussion on upscaling operators).

Before we start interpreting the results of the aggregation of STATSGO data to the catchment scale, we consider it important to discuss some key limitations of STATSGO. Those limitations were already underscored by Miller and White (1998) and also affect more recent soil data sets. It is our impression that although they reduce our ability to correctly reflect soil properties in hydrological models, they are commonly overlooked.

1. Limited depth: Miller and White (1998) note that "[…] only about 2.5% of the components have layers extending below 203 cm (80 inches). Accordingly, the bottom two standard layers contain meaningful data only for a minority of the map units". In other words, although the STATSGO data set is often perceived as describing the top 2.5m of soil, over the majority of the CONUS only the top 1.5m are covered, and data from the bottom 1m in those areas are potentially misleading.

2. Low information content in the deepest layers: They warn the reader that "for approximately half the components, the minimum and maximum depth-to-bedrock […] both have the value 152 cm (60 inches); in the great majority of these cases, this indicates that this was the maximum depth to which soil was normally examined and bedrock was not actually encountered". This means that when the two last layers (1.5 to 2.5m deep) are marked as bed rock, in about half of the cases, the bed rock has not actually been reached, which leads to an underestimation of the soil depth. Given these limitations, we decided to restrict our attention to the top 1.5m of soil (i.e., to the top 9 layers).

3. Only fine fraction characterized: The STATSGO sand, clay and silt fraction are only for the portion of soil that is finer than 2mm. That is, STATSGO data should certainly not be considered to be representative of the whole soil column, but it is also important to keep in mind that it does not either completely characterize its top part, since only the soil fraction finer than 2mm is considered.

4. Lack of representativeness of the dominant soil texture class: Miller and White (1998) stress that STATSGO "units may be quite internally heterogeneous, with as much as 50% of the map unit having soil properties that differ significantly from the map unit description".

5. Scale inadequacy: although soil hydraulic properties can be measured in a lab, it is still unclear how to meaningfully upscale them to the catchment scale (these quantities can be characterized as incommensurate (e.g., Beven, 2012).

In a general sense, soil data sets only characterize the top soil layers, even when the soil can be much deeper (issues 1 and 2). In fact, the soil depth of a catchment indicated as 1.5m in STATSGO can be an order of magnitude greater. Uncertainties in soil depth are critical for hydrological modeling, in particular because they hamper the determination of the root zone storage capacity (Boer-Euser et al., 2016). To explore those uncertainties, we included a recently-released soil data set (Pelletier et al., 2016, refered to as P16 in continutation), from which we extracted the thickness of the permeable layers above bedrock, i.e. the depth to bedrock. The principal advantage of this data set is that it covers the top 50m of soil. It comes on a global 30 arcsec (~1 km) grid. We estimated the catchment average by computing the mean from all the grid points falling within each catchment. However, it does not provide information on soil texture classes, so it cannot be used to estimate variables like the saturated hydraulic conductivity or the porosity. Another key difference is that P16 leveraged geomorphological principles to obtain more precise estimates than what would be obtained by interpolating soil pit observations. We do not explicitly deal with problems 3 to 5 in this study, but expect them to cause lower than expected performance when hydrological modeling relies on STATSGO or similar data sets.

7.2   Results and discussion

Once aggregated to the catchment scale, STATSGO data reveal the following features. Catchments with a sand fraction greater than 50% are predominantly located along the Gulf Coast and the Atlantic Coast, and in the Great Lakes Region (Figure 6a). This leads to a relatively low porosity fraction and high saturated hydraulic conductivity (Figure 6d and e). Conversely, catchments with a silt fraction greater than 50% are mostly located in a band stretching from Kansas to New York (Figure 6b). Catchments in this band also tend to feature a comparatively large clay fraction (Figure 6c). This implies a higher than average porosity fraction (Figure 6d). Although variables like porosity and saturated hydraulic conductivity are commonly relied upon for parameter estimation, we note that their value in terms of process understanding should not be overestimated, given the limitations outlined in Section 7.1.

As for soil depth, STATSGO and P16 both indicate that the soil is shallower in the Appalachian Mountains than along the Gulf and Atlantic Coasts. There are however disagreements in the Rocky Mountains (e.g., in Colorado) and in the Pacific Northwest: STATSGO indicates a soil depth equal or greater than 1.5m when the depth to bed rock according to P16 is smaller than 1m. The lack of quantitative agreement between the two data sets appears clearly in Figure 6i. The left-hand part of the figure (orange background) includes all the catchments in which there can potentially be an agreement between STATSGO and P16, since the estimated depth to bedrock is equal or smaller than 1.5m. There is however considerable scatter around the 1:1 orange curve, which illustrates the uncertainty in estimates of the soil depth, which directly impact estimates of maximum water content of soils (Figure 6f). This issue is even clearer when the right-hand side of Figure 6i is considered. In about half of the catchments (47%), the depth to bed rock is greater than 1.5m, so it cannot be covered by STATSGO. For 24% of the catchments, the depth to bed rock estimated using P16 is greater than 15m, i.e., ten times the depth covered by STATSGO. This underscores the inability of data sets like STATSGO to provide a realistic characterization of soils in areas of high sedimentary deposition.

Finally, we considered three metrics that can be considered as metadata. The fraction of the catchment characterized as "water" is relevant because it indicates the presence of lakes (Figure 6j). The "organic" fraction, which importantly impacts soil hydraulic properties, is typically negligible, but is non-negligible in many catchments in Florida and in the Great Lakes Region (Figure 6k). The fraction of soil marked as "other" (for which no soil characteristics are available, and which is ignored from the computation of all soil attributes) is significant in many catchments (Figure 6l). How detrimental that is will depend on the application. One way to assess this effect when using soil characteristics to explain the performance of hydrological models, would be test whether a clearer relationship is obtained by progressively excluding catchments with the highest fraction of soil marked as "other".

8   Geology

8.1   Data and methods

We used two complementary global sets to characterize the geology of each catchment. The first data set is the Global Lithological Map (GLiM) by Hartmann and Moosdorf (2012). GLiM synthesizes lithological data from 92 regional maps spread across the globe. The spatial resolution is remarkable, as GLiM relies on ~1.2 million polygons to discretize the Earth surface. Three levels of details are available. In this study, we focus on the first level, while the two other levels provide further details that could be processed at a later stage. The first level differentiates between 16 lithological classes (see the list of classes in the legend of Figure 7). We determined the contribution of each lithological classes to the area of each catchment, and recorded the first and second most frequent class within the catchment, as well as the fraction of the catchment they cover. The class "carbonate sedimentary rocks" is particularly relevant from a hydrological perspective (it designates areas likely to host karst systems), we hence also recorded the fraction of each catchment associated with this class. Finally, note that although a 0.5°-gridded version of GLiM is available, we used the more detailed polygon-based version for this study.

The second data set we used to characterize catchment geology is the GLobal HYdrogeology MaPS (GLHYMPS) of the subsurface permeability and porosity by Gleeson et al. (2014). GLHYMPS is based on GLiM spatial polygons, so its level of spatial details is equally high. Gleeson et al. (2014) principally relied on GLiM lithologic classes to derive quantitative estimates of two key characteristics of the geologic units below soil horizons: porosity and permeability (i.e., the ease of fluid flow through porous rocks and soils). For CAMELS, we produced catchment-averages of these two variables, the contribution of each spatial polygon being weighted by the fraction of catchment it covers. The arithmetic mean was used for porosity, but for permeability, we followed Gleeson et al. (2011) and used the geometric mean instead. The geological attributes are summarized in Table 6.

A clear advantage of these high-resolution global lithological maps is that they can be used to extract catchment-scale attributes for diverse parts of the globe. Yet, data quality is spatially variable and caveats of the GLiM and GLYHMPS (outlined in the Section 3 of Gleeson et al., 2014) should be kept in mind. In particular, there are unrealistic spatial discontinuities coinciding with jurisdictional boundaries in GLiM maps, which by construction also affect GLYHMPS maps (for instance in the region of North and South Dakota).

8.2   Results and discussion

The four most frequent dominant geological classes in CAMELS catchments are siliciclastic sedimentary rocks (34% of the catchments), unconsolidated sediments (19%), metamorphics (16%) and carbonate sedimentary rocks (12%). Unconsolidated sediments dominate in catchments along the Gulf Coast and along the southern to middle Atlantic Coast (Figure 7a). In those catchments, both the subsurface porosity (Figure 7f) and permeability (Figure 7g) are high. The Pacific Coast and the region

north of the Appalachian Mountains features catchments rich in siliciclastic sedimentary rocks, leading to a comparatively low subsurface permeability. To the south of the Appalachian Mountains, metamorphic rocks are dominant, resulting in a particularly low subsurface porosity. Finally, the catchments with the highest proportion of carbonate sedimentary rocks are principally located in Central-Western Texas, in the region stretching from Lake Michigan to and including Missouri (Figures 7a and 7f) and to some extent in the Appalachian Mountains (Figure 7b). In addition to these three main regions, there are also isolated catchments with a high proportion of carbonate rocks, for instance in Florida, Nevada and Vermont. The subsurface permeability of those catchments is high. Overall, in 18% of the CAMELS catchments, there is only one GLiM lithological type (Figure 7d), while in 11% of the catchments, the dominant geological class accounts for less than 50% of the catchment area (Figure 7e).

9   Comparison with the MOPEX data set

The CAMELS data set is similar to the data set produced for the Model Parameter Estimation Experiment (MOPEX; Duan et al., 2006; Schaake et al., 2006) in that it provides hydroclimatic time series and geophysical attributes for a large number of basins in the CONUS. MOPEX data have been used in a large number of studies, including two catchment similarity studies mentioned earlier (Sawicz et al., 2011; Berghuijs et al., 2014). For CAMELS, we use different criteria for catchment selection than those used for MOPEX, which leads to a relatively small overlap between the two data sets (they have 52 catchments in common, see Figure 8). The main differences between MOPEX and CAMELS are summarized in Table 7.

Both MOPEX and CAMELS require long observation time series and exclude catchments subject to human influence, but they use different approaches to characterize these aspects. For MOPEX, the stations part of the hydro-climatic data network (HCDN, Slack and Landwehr, 1992) together with those selected by Wallis et al. (1991) were considered to select potential basins. For CAMELS, an updated version of the HCDN classification was used (HCDN-2009, Lins, 2012): some catchments were excluded (e.g., because they no longer met the minimal disturbance criteria defined in the original HCDN report) and other catchments were added (e.g., because their streamflow records, which were considered too short when the original HCDN report was published, became long enough).

For a catchment to be part of the MOPEX data set, an essential criterion was that its number of rain gauges had to be higher than a threshold based on the catchment area. This led to the exclusion of 77% of the potential MOPEX basins, resulting in only 438 basins considered to have a dense enough network of gauges. Although we do recognize the importance of reliable precipitation data for hydrological modeling, we did not exclude catchments based on their rain gage density. We argue that uncertainties in precipitation estimates (and in forcing in general) can now be assessed using independent data sets (e.g., Newman et al., 2015a, see also Section 10 of this paper). We also consider that uncertainties in observed time series (in particular in discharge records, see Coxon et al., 2015; McMillan and Westerberg, 2015) and uncertainties in catchment

attributes (e.g. soil depth, see discussion in Section 7) can also lead to biased conclusions on hydrological processes, and hence should also be considered in the catchment selection processes. Yet the influence of these sources of uncertainties on research results will depend on the catchments and variables of interest, so we leave it to the users to define their own criteria.

5 Overall, the data used for CAMELS are more recent than those used for MOPEX. The period covered by hydro-meteorological times series is 1948-2003 for MOPEX and 1980-2015 for CAMELS, so given the fast rate of human development and the impacts caused by climate change, CAMELS provides a more current picture of hydrological processes in the United States. Further, CAMELS leverages new data sets, which were not available when the MOPEX data were released, for instance to characterize soils (Pelletier et al., 2016) and geology: GLiM (Hartmann and Moosdorf, 2012) and GLHYMPS (Glesson and

10 et al., 2014). And importantly, data used for CAMELS are not only more recent, but also tend to be better documented. A clear example is that CAMELS meteorological time series come from three widely-used gridded data sets (Daymet, Maurer and NLDAS), while for MOPEX, station measurements were aggregated to provide catchment-scale estimates (Schaake et al, 2006).

15 Finally, we present a more detailed and transparent description of the origins and limitations of the data sets used to derive catchment attributes. A substantial part of this paper is dedicated to the discussion of the limitations of the source data sets and we use competing approaches to estimate the same quantity, thereby revealing uncertainties in those attributes. This is motivated by the belief that identifying weaknesses in catchment attributes helps us to anticipate how they might bias conclusions of hydrological studies.

20 **10 Online availability and possible future extensions**

In summary, the CAMELS data set is the combination of two data sets, which are available for download separately:
1. Hydrometeorological time series from Newman et al. (2015b): https://dx.doi.org/10.5065/D6MW2F4D.
2. Catchment attributes introduced in this paper: https://dx.doi.org/10.5065/D6G73C3Q.

25 Our intention with this paper is provide quantitative estimates of key geophysical attributes that shape catchment behavior. We see the data set in its current state as a starting point, and anticipate that it will keep evolving and become more exhaustive. In particular, our next priority is to compute and make available network characteristics and descriptors of catchment geometry, such as drainage density and stream order statistics, which are important for the understanding and simulation of hydrological processes (Rodríguez-Iturbe and Valdés, 1979; Gupta et al., 1980; Rinaldo et al., 1991).

One of our goals is to enable users to assess the reliability of the attributes and to select catchments and interpret results accordingly, and more work is necessary for a complete uncertainty assessment. Methods and data sets are however already available to quantitatively characterize uncertainties in following attributes:

1. Atmospheric forcing: the N15 data set provides forcing from three data sets (Daymet, NLDAS and Maurer) but in this study, we only use Daymet. Using the two other data sets might lead to some differences in climate indices, particularly when it comes to heavy precipitation events and/or to catchments with a sparse observation network. Another option to characterize the uncertainty in the forcing is to use the ensemble of gridded forcing produced by Newman et al. (2015a).

2. Discharge measurements: some hydrological signatures are more sensitive than others to uncertainties in discharge measurements (Westerberg and McMillan, 2015). Methods to characterize those uncertainties and explore their propagation into hydrological signatures in a large sample of catchments exist (Coxon et al., 2015) but require detailed information on the rating curves used for discharge estimation, which were not readily available for this study.

11 Soils: the STATSGO data set is subject to several critical limitations, many of them being overcome by the recently released POLARIS data set (Chaney et al., 2016) and SoilGrids (Hengl et al., 2017). A key advantage of these two data sets is that they describe soil attributes at high resolution, using machine learning algorithms to deliver uncertainty estimates.Concluding remarks

We introduced a new set of attributes for 671 catchments in the contiguous USA. These attributes, together with the hydrometeorolgical time series provided by Newman et al. (2015b) for the same catchments, constitute the CAMELS data set. The wide range of geophysical characteristics covered by these basins opens new opportunities to quantitatively explore how the interplay between topography, climate, land cover, soil and geology shapes hydrological behavior. This enables us to test hypotheses and formulate conclusions valid in diverse conditions and not limited to a few specific locations.

We produced a series of maps depicting catchment attributes over the contiguous USA. We used these maps to examine regional variations of a wide range of attributes and to illustrate the relationships between them. From a practical perspective, our synthesis of several data sources into a single data set at the catchment scale greatly simplifies the comparative study of catchment characteristics and the exploration of their influence on hydrological processes.

An essential feature of this work is that it involves a critical assessment of the limitations of the data and methods used to derive catchment attributes, and a discussion of their consequences for process understanding and hydrological modeling. We highlight in particular uncertainties in soil attributes, and by reviewing the assumptions made during the production and processing of the STATSGO data set, we aim to provide the context necessary to adequately manipulate and interpret these attributes. Other data sets also provide catchment characteristics for a large number of catchments, but deliver them in a deterministic way, without explicitly acknowledging their uncertainties.

Our intention with CAMELS was to start filling this gap. We plan to expand this data set by adding new catchment attributes and refining our characterization of the uncertainties in catchment attributes, forcing and streamflow measurements. We designed the tables of this paper so that they fully describe the methods and data used to compute each attribute, in an effort to make our work transparent and reproducible.

To conclude, we envision that the CAMELS data set will enable progress on a wide range of hydrological challenges related to catchment similarity, model parameter estimation based on geophysical characteristics, model benchmarking, regional variations of model performance, and to the information content of geophysical data sets.

Acknowledgements

10 The National Center for Atmospheric Research (NCAR) is sponsored by the National Science Foundation. This study was supported by the US Army Corps of Engineers Climate Preparedness and Resilience programs. We thank Mike Barlage for his advice on land cover characteristics and Gemma Coxon for her comments on the manuscript. Data processing and visualization was performed with R (R Core Team, 2017). Most colors were chosen using ColorBrewer 2.0 (Brewer, 2017). To conclude, let us highlight that hydrology is all about the storage and transfer of water across the landscape, and so are

15 CAMELS.

References

Abdulla, F. A. and Lettenmaier, D. P.: Development of regional parameter estimation equations for a macroscale hydrologic model, J. Hydrol., 197(1–4), 230–257, doi:10.1016/S0022-1694(96)03262-3, 1997.

Addor, N., Rössler, O., Köplin, N., Huss, M., Weingartner, R. and Seibert, J.: Robust changes and sources of uncertainty in

20 the projected hydrological regimes of Swiss catchments, Water Resour. Res., 50(10), 7541–7562, doi:10.1002/2014WR015549, 2014.

Aguado, E., Cayan, D., Riddle, L. and Roos, M.: Climatic Fluctuations and the Timing of West Coast Streamflow, J. Clim., 5, 1468–1483, doi:10.1175/1520-0442(1992)005<1468:CFATTO>2.0.CO;2, 1992.

Beck, H. E., Van Dijk, A. I. J. M., Miralles, D. G., De Jeu, R. A. M., Bruijnzeel, L. A., McVicar, T. R. and Schellekens, J.:

25 Global patterns in base flow index and recession based on streamflow observations from 3394 catchments, Water Resour. Res., 49(12), 7843–7863, doi:10.1002/2013WR013918, 2013.

Beck, H. E., van Dijk, A. I. J. M., Roo, A. de, Miralles, D. G., McVicar, T. R., Schellekens, J. and Bruijnzeel, L. A.: Global-scale regionalization of hydrologic model parameters, Water Resour. Res., 52(5), 3599–3622, doi:10.1002/2015WR018247, 2016.

30 Belward, A. E.: The IGBP-DIS global 1 km land cover data set "DISCover" - Proposal and implementation plans. Report of

the Land Cover Working Group of the IGBP-DIS. IGBP-DIS Working Paper, No. 13, Stockholm., 1996.

Berghuijs, W. R., Sivapalan, M., Woods, R. A. and Savenije, H. H. G.: Patterns of similarity of seasonal water balances: A window into streamflow variability over a range of time scales, Water Resour. Res., 50, 5638–5661, doi:10.1002/2014WR015692, 2014.

5 Berghuijs, W. R. and Woods, R. A.: A simple framework to quantitatively describe monthly precipitation and temperature climatology, Int. J. Climatol., 36(9), 3161–3174, doi:10.1002/joc.4544, 2015.

Beven, K. J.: Rainfall-Runoff Modelling: The Primer, 2nd Edition, Wiley-Blackwell., 2012.

Bock, A. R., Hay, L. E., McCabe, G. J., Markstrom, S. L. and Atkinson, R. D.: Parameter regionalization of a monthly water balance model for the conterminous United States, Hydrol. Earth Syst. Sci., 20(7), 2861–2876, doi:10.5194/hess-20-2861-

10 2016, 2016.

Boer-Euser, T. de, McMillan, H. K., Hrachowitz, M., Winsemius, H. C. and Savenije, H. H. G.: Influence of soil and climate on root zone storage capacity, Water Resour. Res., 52, 2009–2024, doi:10.1002/2015WR018115, 2016.

Brewer, C. A.: ColorBrewer, [online] Available from: http://colorbrewer2.org/, 2017.

Chaney, N. W., Wood, E. F., McBratney, A. B., Hempel, J. W., Nauman, T. W., Brungard, C. W. and Odgers, N. P.: POLARIS:

15 A 30-meter probabilistic soil series map of the contiguous United States, Geoderma, 274, 54–67, doi:10.1016/j.geoderma.2016.03.025, 2016.

Clark, M. P., McMillan, H. K., Collins, D. B. G., Kavetski, D. and Woods, R. A.: Hydrological field data from a modeller's perspective: Part 2: Process-based evaluation of model hypotheses, Hydrol. Process., 25(4), 523–543, doi:10.1002/hyp.7902, 2011.

20 Clausen, B. and Biggs, B. J. F.: Flow variables for ecological studies in temperate streams: Groupings based on covariance, J. Hydrol., 237(3–4), 184–197, doi:10.1016/S0022-1694(00)00306-1, 2000.

Cosby, B. J., Hornberger, G. M., Clapp, R. B. and Ginn, T. R.: A Statistical Exploration of the Relationships of Soil Moisture Characteristics to the Physical Properties of Soils, Water Resour. Res., 20(6), 682–690, doi:10.1029/WR020i006p00682, 1984.

Court, A.: Measures of streamflow timing, J. Geophys. Res., 67(11), 4335–4339, doi:10.1029/JZ067i011p04335, 1962.

25 Coxon, G., Freer, J., Westerberg, I. K., Wagener, T., Woods, R. and Smith, P. J.: A novel framework for discharge uncertainty quantification applied to 500 UK gauging stations, Water Resour. Res., 51, 5531–5546, doi:10.1002/2014WR016259, 2015.

Duan, Q., Schaake, J., Andréassian, V., Franks, S., Goteti, G., Gupta, H. V., Gusev, Y. M., Habets, F., Hall, A., Hay, L., Hogue, T., Huang, M., Leavesley, G., Liang, X., Nasonova, O. N., Noilhan, J., Oudin, L., Sorooshian, S., Wagener, T. and Wood, E. F.: Model Parameter Estimation Experiment (MOPEX): An overview of science strategy and major results from the

30 second and third workshops, J. Hydrol., 320(1–2), 3–17, doi:10.1016/j.jhydrol.2005.07.031, 2006.

Falcone, J. A.: GAGES-II: Geospatial Attributes of Gages for Evaluating Streamflow. [online] Available from: https://water.usgs.gov/GIS/metadata/usgswrd/XML/gagesII_Sept2011.xml, 2011.

Falkenmark, M. and Chapman, T.: Comparative hydrology An ecological approach to land and water resources Unesco, UNESCO, Paris., 1989.

Fenicia, F., Kavetski, D., Savenije, H. H. G., Clark, M. P., Schoups, G., Pfister, L. and Freer, J.: Catchment properties, function, and conceptual model representation: is there a correspondence?, Hydrol. Process., 28(4), 2451–2467, doi:10.1002/hyp.9726, 2014.

Gleeson, T., Moosdorf, N., Hartmann, J. and van Beek, L. P. H.: A glimpse beneath earth's surface: GLobal HYdrogeology

5   MaPS (GLHYMPS) of permeability and porosity, Geophys. Res. Lett., 41, 3891–3898, doi:10.1002/2014GL059856, 2014.

Gleeson, T., Smith, L., Moosdorf, N., Hartmann, J., Dürr, H. H., Manning, A. H., Van Beek, L. P. H. and Jellinek, A. M.: Mapping permeability over the surface of the Earth, Geophys. Res. Lett., 38(2), 1–6, doi:10.1029/2010GL045565, 2011.

Gupta, H. V., Perrin, C., Blöschl, G., Montanari, A., Kumar, R., Clark, M. and Andréassian, V.: Large-sample hydrology: a need to balance depth with breadth, Hydrol. Earth Syst. Sci., 18(2), 463–477, doi:10.5194/hess-18-463-2014, 2014.

10   Gupta, V. K., Waymire, E. and Wang, C. T.: A representation of an instantaneous unit hydrograph from geomorphology, Water Resour. Res., 16(5), 855–862, doi:10.1029/WR016i005p00855, 1980.

Harman, C. J., Troch, P. A. and Sivapalan, M.: Functional model of water balance variability at the catchment scale: 2. Elasticity of fast and slow runoff components to precipitation change in the continental United States, Water Resour. Res., 47(2), 1–12, doi:10.1029/2010WR009656, 2011.

15   Harman, C. and Troch, P. A.: What makes Darwinian hydrology "darwinian"? Asking a different kind of question about landscapes, Hydrol. Earth Syst. Sci., 18(2), 417–433, doi:10.5194/hess-18-417-2014, 2014.

Hartmann, J. and Moosdorf, N.: The new global lithological map database GLiM: A representation of rock properties at the Earth surface, Geochemistry, Geophys. Geosystems, 13(12), 1–37, doi:10.1029/2012GC004370, 2012.

Hengl, T., Mendes De Jesus, J., Heuvelink, G. B. M., Gonzalez, M. R., Kilibarda, M., Blagotí, A., Shangguan, W., Wright, M.

20   N., Geng, X., Bauer-Marschallinger, B., Guevara, M. A., Vargas, R., MacMillan, R. A., Batjes, N. H., Leenaars, J. G. B., Ribeiro, E., Wheeler, I., Mantel, S. and Kempen, B.: SoilGrids250m: Global Gridded Soil Information Based on Machine Learning., 2017.

Hrachowitz, M., Savenije, H. H. G., Blöschl, G., McDonnell, J. J., Sivapalan, M., Pomeroy, J. W., Arheimer, B., Blume, T., Clark, M. P., Ehret, U., Fenicia, F., Freer, J. E., Gelfan, A., Gupta, H. V., Hughes, D. a., Hut, R. W., Montanari, A., Pande, S.,

25   Tetzlaff, D., Troch, P. a., Uhlenbrook, S., Wagener, T., Winsemius, H. C., Woods, R. a., Zehe, E. and Cudennec, C.: A decade of Predictions in Ungauged Basins (PUB)—a review, Hydrol. Sci. J., 58(6), 1198–1255, doi:10.1080/02626667.2013.803183, 2013.

Hundecha, Y., Ouarda, T. B. M. J. and Bárdossy, A.: Regional estimation of parameters of a rainfall-runoff model at ungauged watersheds using the "spatial" structures of the parameters within a canonical physiographic-climatic space, Water Resour.

30   Res., 44(1), 1–13, doi:10.1029/2006WR005439, 2008.

Ladson, A., Brown, R., Neal, B. and Nathan, R.: A standard approach to baseflow separation using the Lyne and Hollick filter, Aust. J. Water Resour., 17(1), 25–34, doi:http://dx.doi.org/10.7158/W12-028.2013.17.1., 2013.

Lawrence, D. M. and Slater, A. G.: Incorporating organic soil into a global climate model, Clim. Dyn., 30(2–3), 145–160, doi:10.1007/s00382-007-0278-1, 2008.

Lins, H.: USGS hydro-climatic data network 2009 (HCDN-2009)., 2012.

Maurer, E. P., Wood, A. W., Adam, J. C., Lettenmaier, D. P. and Nijssen, B.: A Long-Term Hydrologically Based Dataset of Land Surface Fluxes and States for the Conterminous United States, J. Clim., 26(23), 9384–9392, doi:10.1175/JCLI-D-12-00508.1, 2002.

5   McDonnell, J. J. and Woods, R.: On the need for catchment classification, J. Hydrol., 299(1–2), 2–3, doi:10.1016/j.jhydrol.2004.09.003, 2004.

McMillan, H. K., Clark, M. P., Bowden, W. B., Duncan, M. and Woods, R. A.: Hydrological field data from a modeller's perspective: Part 1. Diagnostic tests for model structure, Hydrol. Process., 25(4), 511–522, doi:10.1002/hyp.7841, 2011.

McMillan, H. K. and Westerberg, I. K.: Rating curve estimation under epistemic uncertainty, Hydrol. Process., 29(7), 1873–

10   1882, doi:10.1002/hyp.10419, 2015.

Miller, D. A. and White, R. A.: A Conterminous United States Multilayer Soil Characteristics Dataset for Regional Climate and Hydrology Modeling, Earth Interact., 2, doi:10.1175/1087-3562(1998)002<0002:CUSMS>2.0.CO;2, 1998.

Newman, A. J., Clark, M. P., Craig, J., Nijssen, B., Wood, A., Gutmann, E., Mizukami, N., Brekke, L. and Arnold, J. R.: Gridded Ensemble Precipitation and Temperature Estimates for the Contiguous United States, J. Hydrometeorol., 16(6),

15   150817120004007, doi:10.1175/JHM-D-15-0026.1, 2015a.

Newman, A. J., Clark, M. P., Sampson, K., Wood, A., Hay, L. E., Bock, A., Viger, R., Blodgett, D., Brekke, L., Arnold, J. R., Hopson, T. and Duan, Q.: Development of a large-sample watershed-scale hydrometeorological dataset for the contiguous USA: dataset characteristics and assessment of regional variability in hydrologic model performance, Hydrol. Earth Syst. Sci., 19, 209–223, doi:10.5194/hess-19-209-2015, 2015b.

20   Olden, J. D. and Poff, N. L.: Redundancy and the choice of hydrologic indices for characterizing streamflow regimes, River Res. Appl., 19(2), 101–121, doi:10.1002/rra.700, 2003.

Oudin, L., Andréassian, V., Lerat, J. and Michel, C.: Has land cover a significant impact on mean annual streamflow? An international assessment using 1508 catchments, J. Hydrol., 357(3–4), 303–316, doi:10.1016/j.jhydrol.2008.05.021, 2008.

Pelletier, J. D., Patrick D. Broxton, Hazenberg, P., Zeng, X., Troch, P. A., Niu, G.-Y., Williams, Z., Brunke, M. A. and Gochis,

25   D.: A gridded global data set of soil, intact regolith, and sedimentary deposit thicknesses for regional and global land surface modeling, J. Adv. Model. Earth Syst., doi:10.1002/2013MS000282.Received, 2016.

R Core Team: R: A Language and Environment for Statistical Computing, [online] Available from: http://www.r-project.org/, 2017.

Rinaldo, A., Marani, A. and Rigon, R.: Geomorphological dispersion, Water Resour. Res., 27(4), 513–525,

30   doi:10.1029/90WR02501, 1991.

Rodríguez-Iturbe, I. and Valdés, J. B.: Geomorphologic Structure of Hydrologic Response, Water Resour. Res., 15(6), 1409–1420, doi:10.1029/WR015i006p01409, 1979.

Samaniego, L., Kumar, R. and Attinger, S.: Multiscale parameter regionalization of a grid-based hydrologic model at the mesoscale, Water Resour. Res., 46(5), doi:10.1029/2008WR007327, 2010.

Sawicz, K., Wagener, T., Sivapalan, M., Troch, P. a. and Carrillo, G.: Catchment classification: Empirical analysis of hydrologic similarity based on catchment function in the eastern USA, Hydrol. Earth Syst. Sci., 15(9), 2895–2911, doi:10.5194/hess-15-2895-2011, 2011.

Schaake, J., Cong, S. Z. and Duan, Q. Y.: The US MOPEX data set., IAHS Publ. 307, (c), 9–28, 2006.

Seibert, J.: Regionalisation of parameters for a conceptual rainfall-runoff model, Agric. For. Meteorol., 98–99, 279–293, doi:10.1016/S0168-1923(99)00105-7, 1999.

Seybold, H., Rothman, D. H. and Kirchner, J. W.: Climate's watermark in the geometry of stream networks, Geophys. Res. Lett., doi:10.1002/2016GL072089, 2017.

Slack, J. R. and Landwehr, J. M.: Hydro-climatic data network (HCDN); a U.S. Geological Survey streamflow data set for the United States for the study of climate variations, 1874-1988., 1992.

Stewart, I., Cayan, D. and Dettinger, M.: Changes toward earlier streamflow timing across western North America., J. Clim., 1136–1155, doi:10.1175/JCLI3321.1, 2005.

Thompson, S. E., Harman, C. J., Konings, A. G., Sivapalan, M., Neal, A. and Troch, P. A.: Comparative hydrology across AmeriFlux sites: The variable roles of climate, vegetation, and groundwater, Water Resour. Res., 47(7), 1–17, doi:10.1029/2010WR009797, 2011.

Thornton, P. E., Thornton, M. M., Mayer, B. W., Wilhelmi, N., Wei, Y. and Cook, R. B.: Daymet: Daily surface weather on a 1km grid for North America, 1980–2012., 2012.

Troch, P. A., Martinez, G. F., Pauwels, V. R. N., Durcik, M., Sivapalan, M., Harman, C., Brooks, P. D., Gupta, H. and Huxman, T.: Climate and vegetation water use efficiency at catchment scales, Hydrol. Process., 23, 2409–2414, doi:10.1002/hyp.7358, 2009.

Vano, J. A., Nijssen, B. and Lettenmaier, D. P.: Seasonal hydrologic responses to climate change in the Pacific Northwest, Water Resour. Res., 51, doi:10.1002/2014WR015909, 2015.

Viger, R. J.: Preliminary spatial parameters for PRMS based on the Geospatial Fabric, NLCD2001 and SSURGO., 2014.

Viger, R. J. and Bock, A.: GIS Features of the Geospatial Fabric for National Hydrologic Modeling., 2014.

Voepel, H., Ruddell, B., Schumer, R., Troch, P. A., Brooks, P. D., Neal, A., Durcik, M. and Sivapalan, M.: Quantifying the role of climate and landscape characteristics on hydrologic partitioning and vegetation response, Water Resour. Res., 47(8), 1–13, doi:10.1029/2010WR009944, 2011.

Wagener, T., Sivapalan, M., Troch, P. and Woods, R.: Catchment Classification and Hydrologic Similarity, Geogr. Compass, 1, 1–31, doi:10.1111/j.1749-8198.2007.00039.x, 2007.

Wallis, J. R., Lettenmaier, D. P. and Wood, E. F.: A daily hydroclimatological data set for the continental United States, Water Resour. Res., 27(7), 1657–1663, doi:10.1029/91WR00977, 1991.

Westerberg, I. K. and McMillan, H. K.: Uncertainty in hydrological signatures, Hydrol. Earth Syst. Sci., 19(9), 3951–3968, doi:10.5194/hess-19-3951-2015, 2015.

Westerberg, I. K., Wagener, T., Coxon, G., McMillan, H. K., Castellarin, A., Montanari, A. and Freer, J.: Uncertainty in

hydrological signatures for gauged and ungauged catchments, Water Resour. Res., 52, 1847–1865, doi:10.1002/2015WR017635, 2016.

Xia, Y., Mitchell, K., Ek, M., Sheffield, J., Cosgrove, B., Wood, E., Luo, L., Alonge, C., Wei, H., Meng, J., Livneh, B., Lettenmaier, D., Koren, V., Duan, Q., Mo, K., Fan, Y. and Mocko, D.: Continental-scale water and energy flux analysis and

5   validation for the North American Land Data Assimilation System project phase 2 (NLDAS-2): 1. Intercomparison and application of model products, J. Geophys. Res., 117(D3), D03109, doi:10.1029/2011jd016048, 2012.

Zeng, X.: Global Vegetation Root Distribution for Land Modeling, J. Hydrometeorol., 2(5), 525–530, doi:10.1175/1525-7541(2001)002<0525:GVRDFL>2.0.CO;2, 2001.

Zhang, L., Dawes, W. R. and Walker, G. R.: Response of mean annual evapotranspiration to vegetation changes at catchment

10   scale, Water Resour. Res., 37, 701–708, doi:10.1029/2000WR900325, 2001.
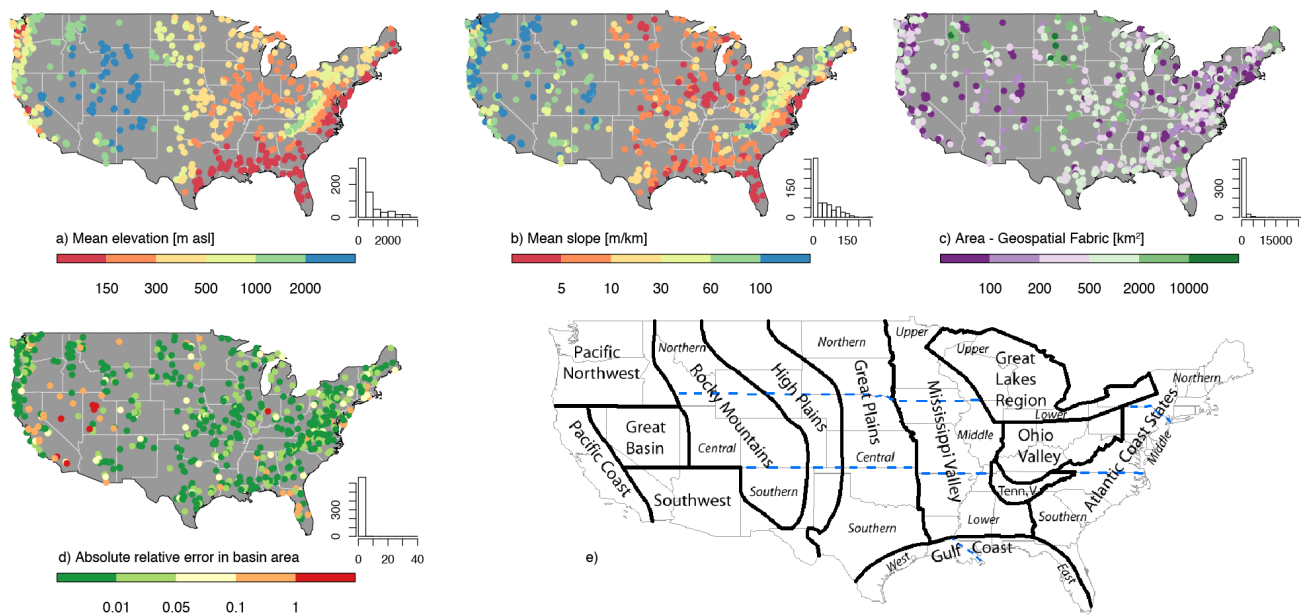
Figure 1: a-d) Maps of topographic variables over the CONUS. The histograms indicate the number of catchments (out of 671) in each bin. e) Map of the regions referred to in this study (source: NOAA National Centers for Environmental Information, https://www.ncdc.noaa.gov/temp-and-precip/drought/nadm/geography)
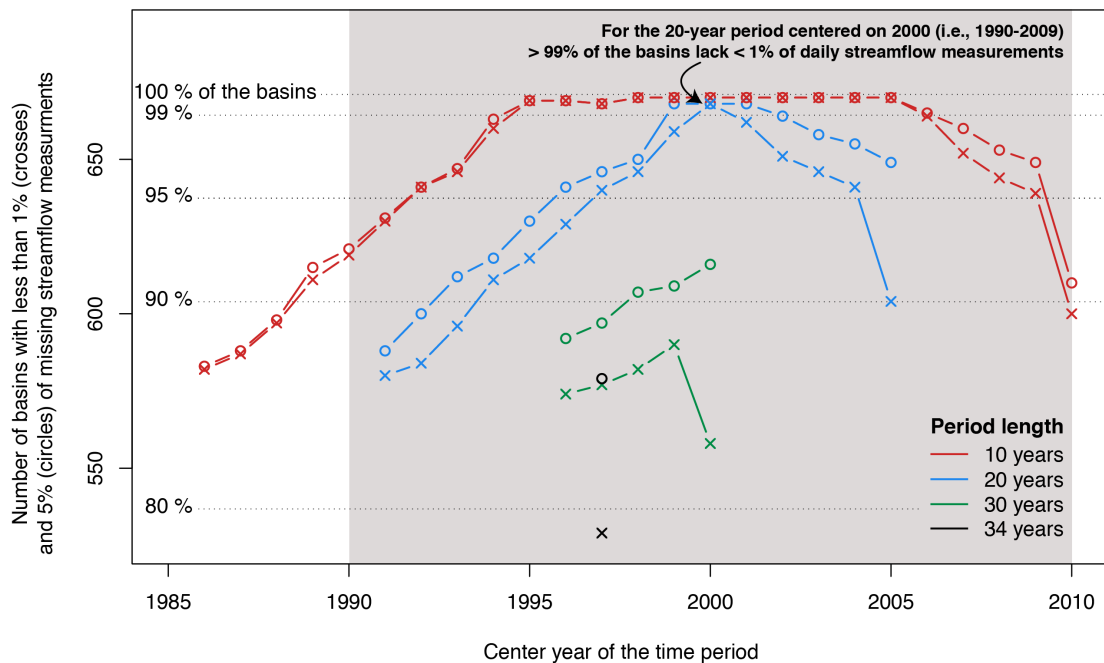
Figure 2: Availability of streamflow measurements for periods of different length (colors) centered on different years (x-axis). The symbols (crosses and circles) indicate the number of catchments with at most 1% or 5% of daily streamflow measurements missing, respectively. The shape of the curves indicates that the proportion of missing data decreases from 1980 to 1990, stays low and then increases after 2010. Note that years are hydrological years (starting on October 1ˢᵗ).
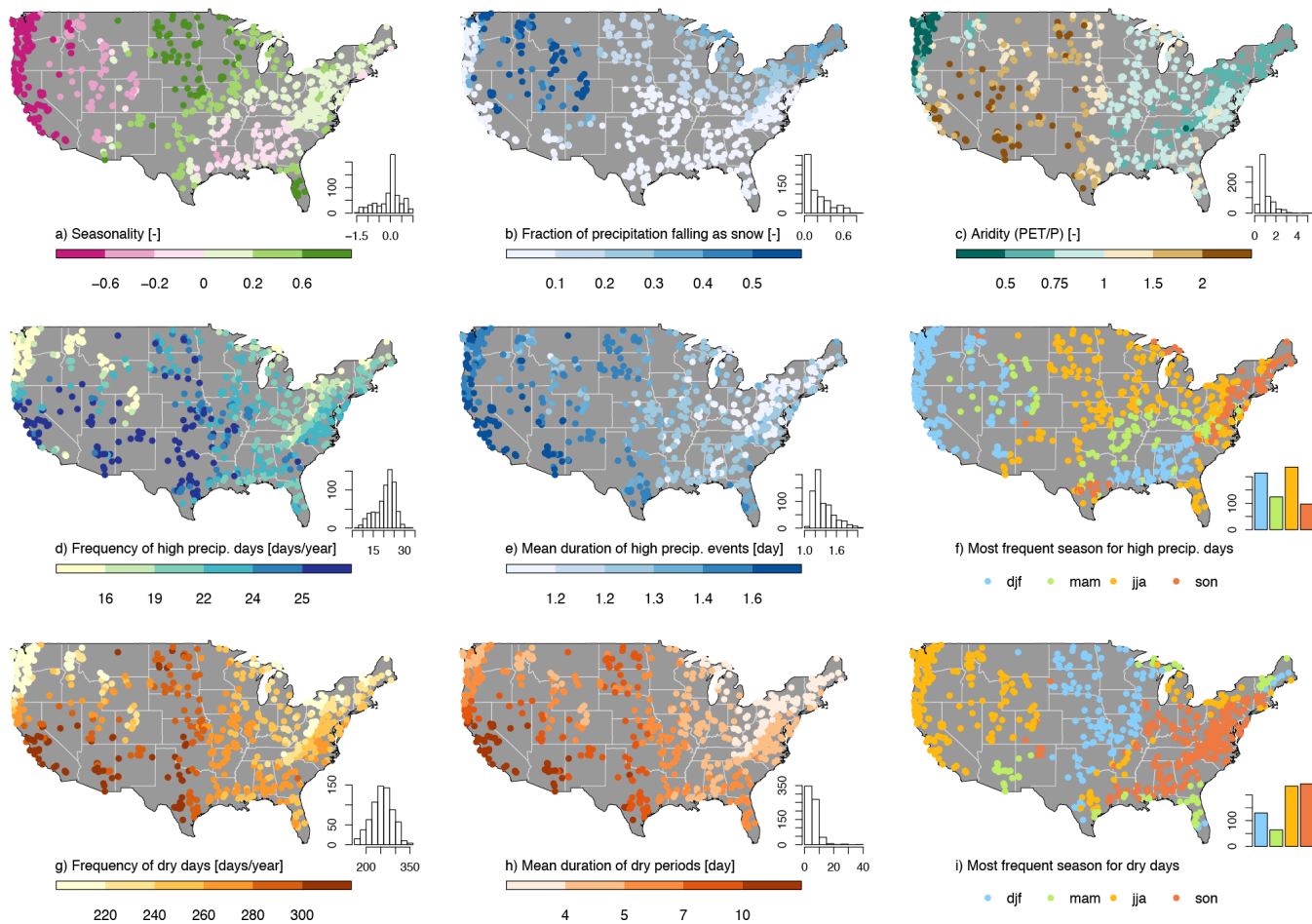
Figure 3: Maps of the climate indices over the CONUS. The histograms and barplots indicate the number of catchments (out of 671) in each bin or category.
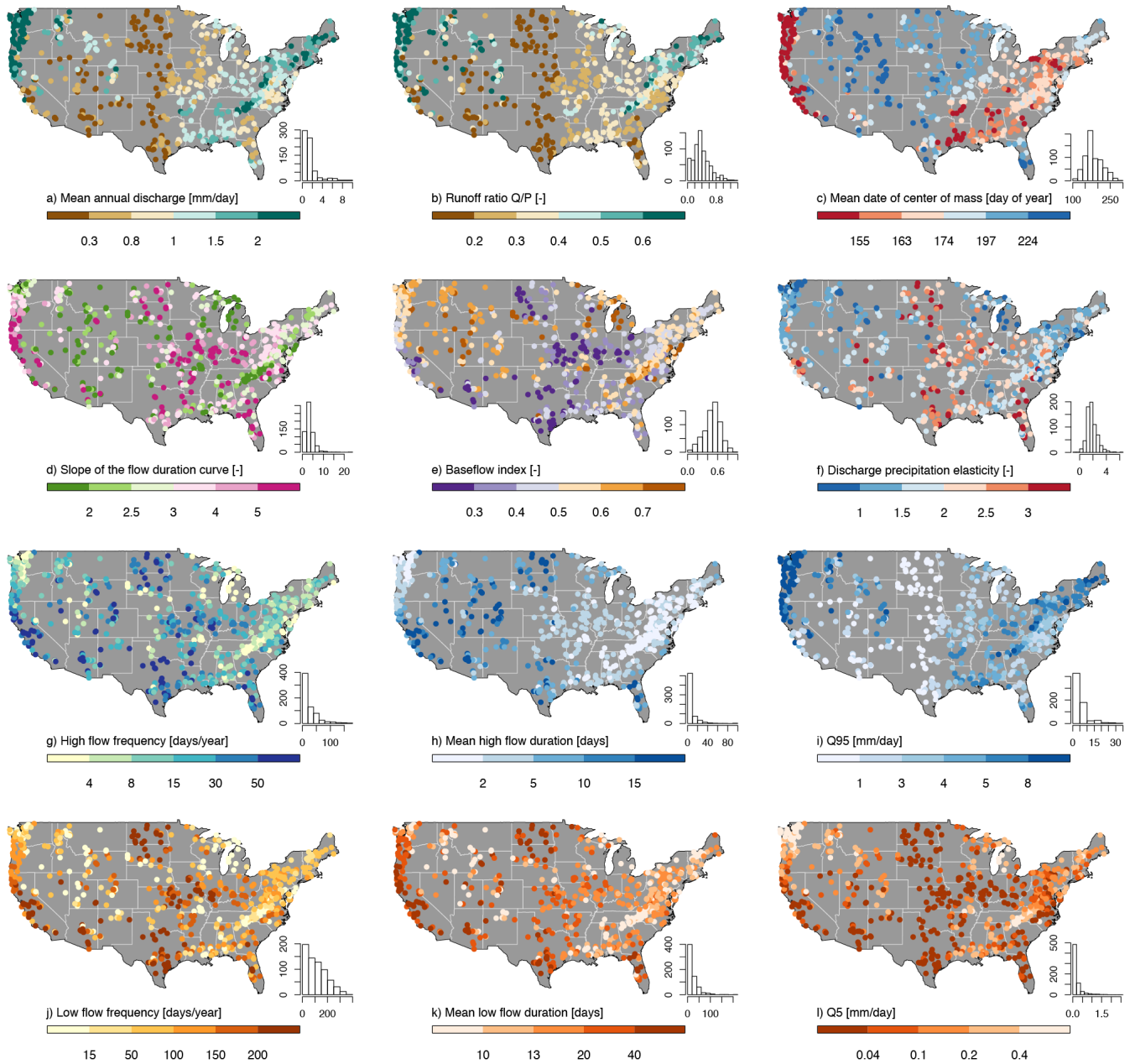
Figure 4: Maps of the hydrological signatures over the CONUS. The histograms indicate the number of catchments (out of 671) in each bin.
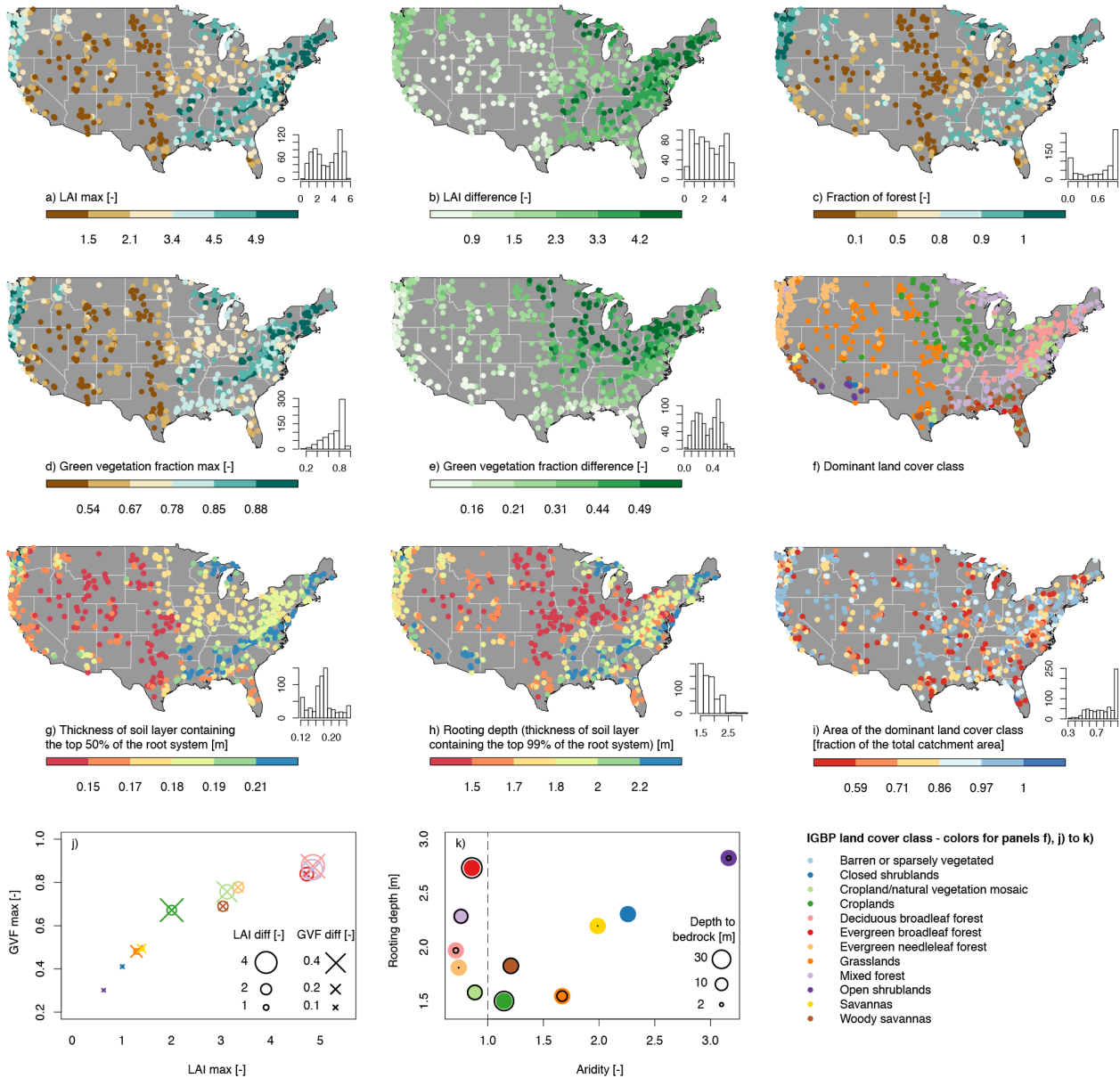
Figure 5: a) to i) Maps of the vegetation characteristics over the CONUS. The histograms indicate the number of catchments (out of 671) in each bin. j) Comparison of the LAI and GVF (maximum and difference between maximum and minimum) based on the dominant land cover class. k) Comparison of the mean aridity, mean total rooting depth and depth to bedrock for each land cover class, the color dots all have the same size.
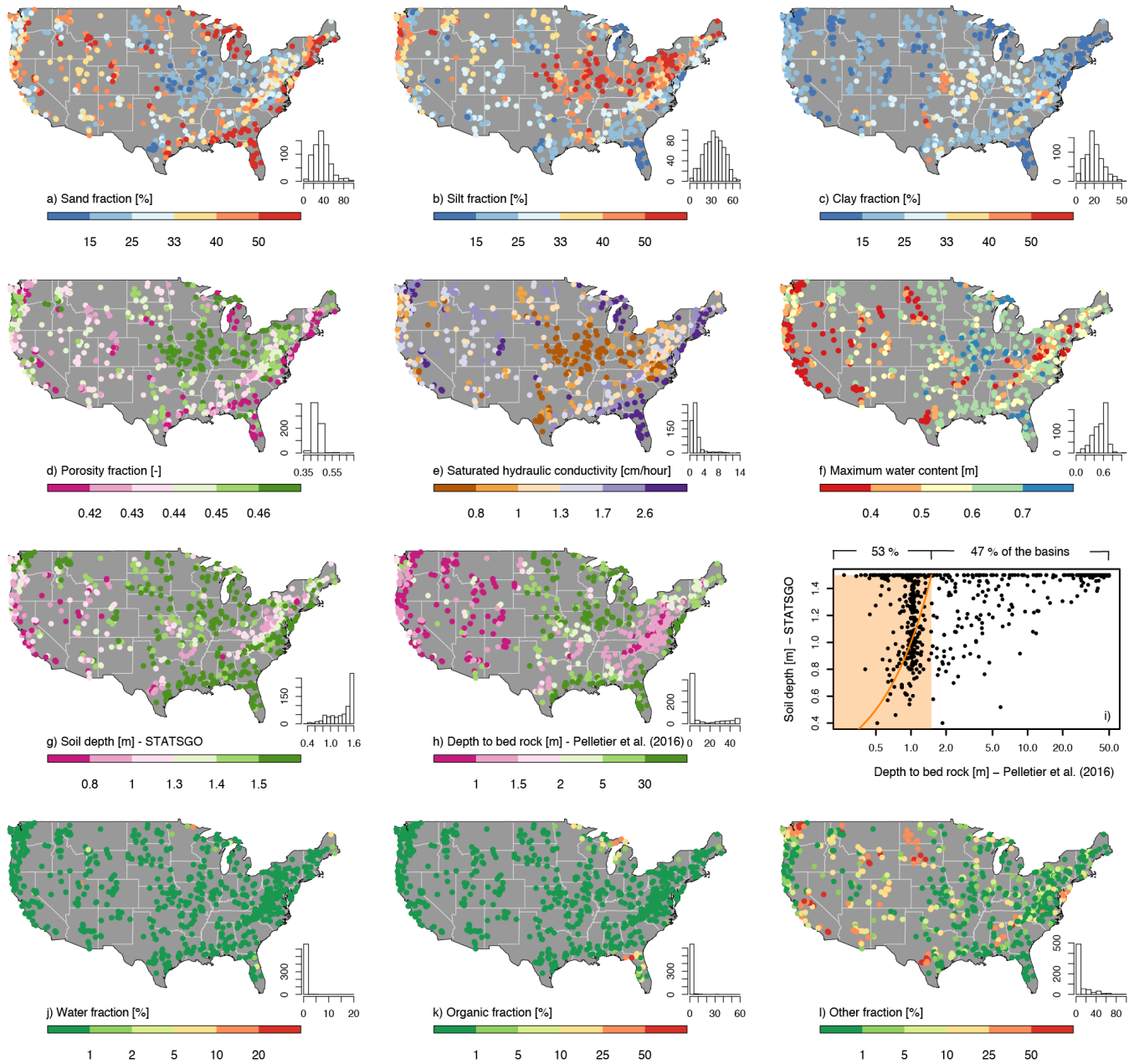
Figure 6: a-h and j-l) Maps of the soil characteristics over the CONUS. The histograms indicate the number of catchments (out of 671) in each bin. **i)** Comparison of the estimates of the depth to bed rock from STATSGO and P16. The orange area includes all the catchments for which there can be an agreement between the two data sets (i.e. estimates from both data sets are smaller than or equal to 1.5m). The orange curve is a 1:1 curve (note the logarithmic scale on the x-axis).
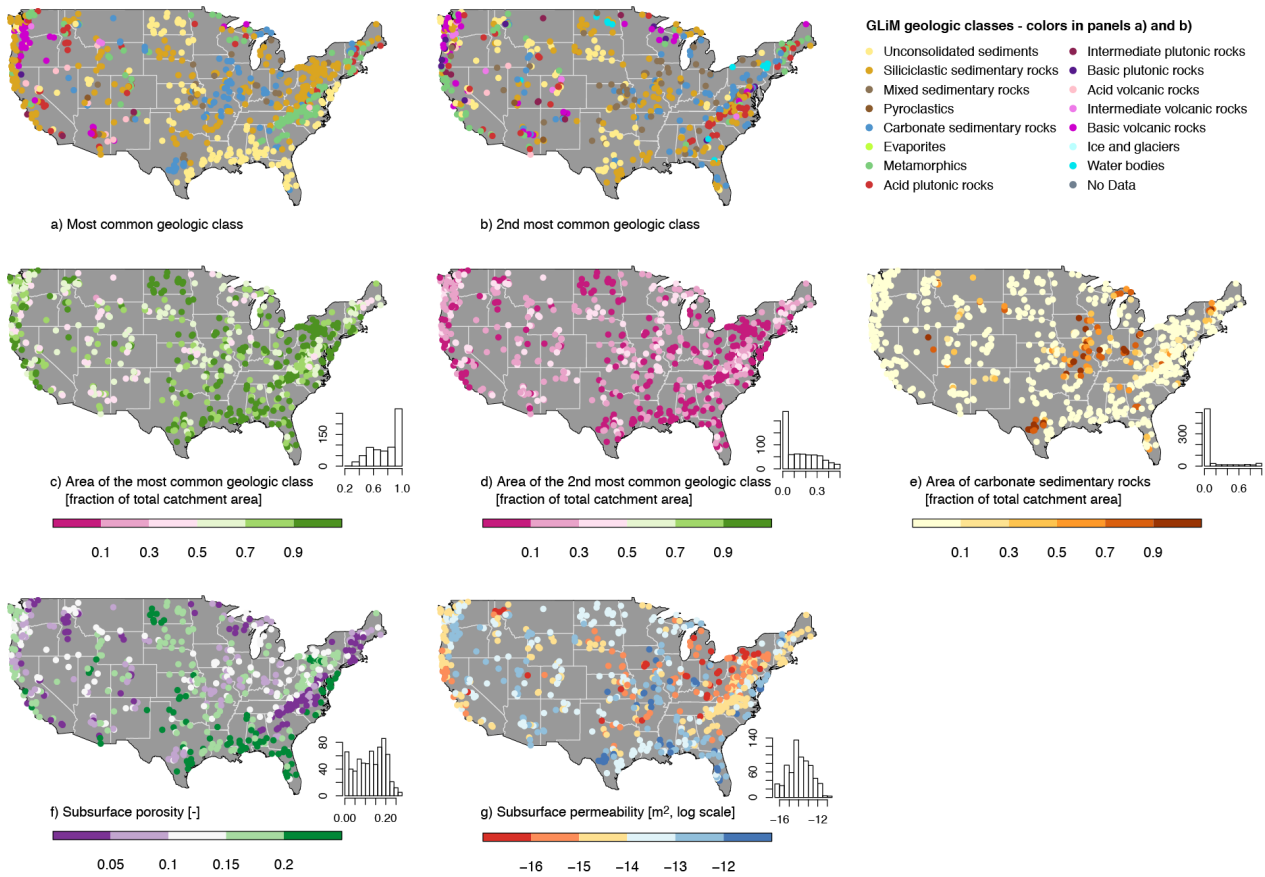
Figure 7: Maps of the geological characteristics over the CONUS. The histograms indicate the number of catchments (out of 671) in each bin.
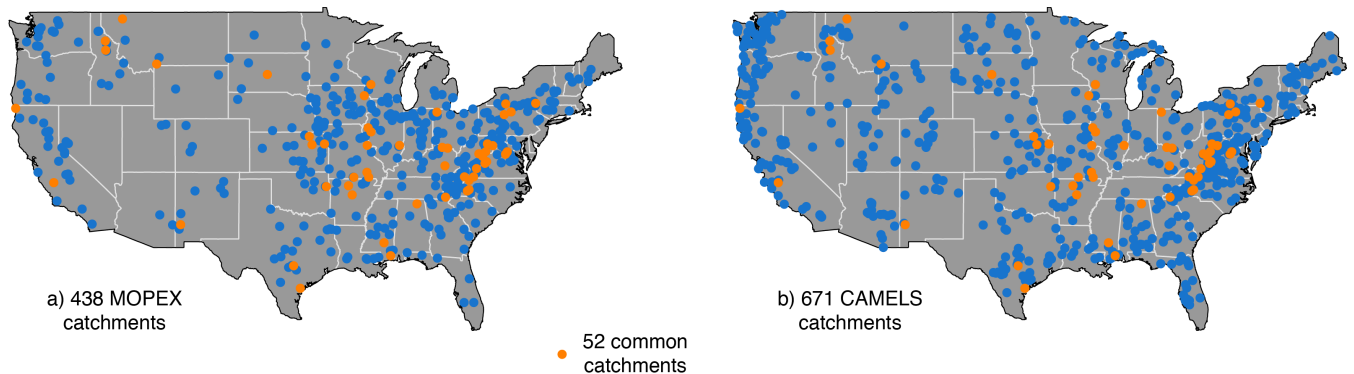
Figure 8: A comparison of the spatial distribution of the catchments from a) the MOPEX data set and b) the CAMELS data set.

Table 1: Name and topography

Attribute	Description	Unit	Data source	References
gauge_id	catchment identifier (8-digit USGS hydrologic unit code)	-	N15 - USGS data	-
huc_02	region (2-digit USGS hydrologic unit code)	-	N15 - USGS data	-
gauge_name	gauge name, followed by the state	-	N15 - USGS data	-
gauge_lat	gauge latitude	° north	N15 - USGS data	-
gauge_lon	gauge longitude	° east	N15 - USGS data	-
elev_mean	catchment mean elevation	meter above sea level	N15 - USGS data	-
slope_mean	catchment mean slope	m/km	N15 - USGS data	-
area_gages2	catchment area (GAGESII estimate)	km^2	N15 - USGS data	Falcone (2011)
area_geospa_fabric	catchment area (Geospatial Fabric estimate)	km^2	N15 - Geospatial Fabric	Viger (2014), Viger and Bock (2014)

Table 2: Climate indices. *: Computed over the period 1989/10/01 to 2009/09/30

Attribute	Description	Unit	Data source	References
p_mean	mean daily precipitation	mm/day	N15 - Daymet*	
pet_mean	mean daily PET [estimated by N15 using Priestley-Taylor formulation calibrated for each catchment]	mm/day	N15 - Daymet*	
aridity	aridity (PET/P, ratio of mean PET [estimated by N15 using Priestley-Taylor formulation calibrated for each catchment] to mean precipitation)	-	N15 - Daymet*	-
p_seasonality	seasonality and timing of precipitation (estimated using sine curves to represent the annual temperature and preciptiation cycles, positive [negative] values indicate that precipitation peaks in summer [winter], values close to 0 indicate uniform precipitation throughout the year)	-	N15 - Daymet*	Eq. 14 in Woods et al. (2009)
frac_snow_daily	fraction of precipitation falling as snow (i.e. on days colder than 0°C)	-	N15 - Daymet*	-
high_prec_freq	frequency of high precipitation days (>= 5 times mean daily precip.)	days/year	N15 - Daymet*	-
high_prec_dur	average duration of high precipitation events (number of consecutive days >= 5 times mean daily precip.)	days	N15 - Daymet*	-
high_prec_timing	season during which most high precipitation days (>= 5 times mean daily precip.) occur	qualitative	N15 - Daymet*	-
low_prec_freq	average frequency of dry days (<1 mm/day)	days/year	N15 - Daymet*	-
low_prec_dur	average duration of dry periods (number of consecutive days <1 mm/day)	days	N15 - Daymet*	-
low_prec_timing	season during which most dry days (<1 mm/day) occur	qualitative	N15 - Daymet*	-

Table 3: Hydrological signatures. *: Computed over the period 1989/10/01 to 2009/09/30

Attribute	Description	Unit	Data source	References
q_mean	mean discharge	mm/day	N15 - USGS data*	-
runoff_ratio	runoff ratio (ratio of long-term average streamflow to long-term average precipitation)	-	N15 - USGS data*	Eq. 2 in Sawicz et al. (2011)
stream_elas	streamflow elasticity (sensitivity of streamflow to changes in precipitation at the annual time scale, using the long-term average as reference)	-	N15 - USGS data*	Eq. 7 in Sankarasubramanian et al. (2011), the last element being P/Q not Q/P
slope_fdc	slope of the flow duration curve (between the log-transformed 33rd and 66th streamflow percentiles)		N15 - USGS data*	Eq. 3 in Sawicz et al. (2011)
baseflow_index	baseflow index (ratio of long-term average baseflow to long-term average streamflow, hydrograph separation performed using Ladson et al. [2013] digital filter)	-	N15 - USGS data*	Eq. 4 in Sawicz et al. (2011), Ladson et al. (2013)
hfd_mean	mean half flow date (date on which the cumulative discharge since October 1st reaches half of the annual discharge)	day of year	N15 - USGS data*	Court (1962)
q5	5% flow quantile	mm/day	N15 - USGS data*	-
q95	95% flow quantile	mm/day	N15 - USGS data*	-
high_q_freq	average frequency of high flow days (> 9 times the median daily flow)	days/year	N15 - USGS data*	Clausen and Biggs (2000), Table 2 in Westerberg and McMillan (2015)
high_q_dur	average duration of high flow events (number of consecutive days > 9 times the median daily flow)	days	N15 - USGS data*	Clausen and Biggs (2000), Table 2 in Westerberg and McMillan (2015)
low_q_freq	average frequency of daily low-flow days (< 0.2 times mean daily flow)	days/year	N15 - USGS data*	Olden and Poff (2003), Table 2 in Westerberg and McMillan (2015)
low_q_dur	average duration of daily flow events (number of consecutive days < 0.2 times the mean daily flow)	days	N15 - USGS data*	Olden and Poff (2003), Table 2 in Westerberg and McMillan (2015)
zero_q_freq	frequency of days with Q = 0	%	N15 - USGS data*	

Table 4: Land cover characteristics. *: Computed over the period 2002 to 2014.

Attribute	Description	Unit	Data source	References
forest_frac	forest fraction	-	N15 - USGS data	-
lai_max	maximum monthly mean of the leaf area index (based on 12 monthly means)	-	MODIS*	-
lai_diff	difference between the maximum and mimumum monthly mean of the leaf area index (based on 12 monthly means)	-	MODIS*	-
gvf_max	maximum monthly mean of the green vegetation fraction (based on 12 monthly means)	-	MODIS*	-
gvf_diff	difference between the maximum and mimumum monthly mean of the green vegetation fraction (based on 12 monthly means)	-	MODIS*	-
dom_land_cover	dominant land cover (Noah-modified 20-category IGBP-MODIS land cover)	-	MODIS*	-
dom_land_cover_frac	fraction of the catchment area associated with the dominant land cover	-	MODIS*	-
root_depth_XX	root depth (percentiles XX = 50 and 99% extracted from a root depth distribution based on IGBP land cover)	m	MODIS*	Eq. 2 and Table 2 in Zeng (2011)

Table 5: Soil characteristics. *: Only covers the top 1.5m.

Attribute	Description	Unit	Data source	References
soil_depth_pelletier	depth to bedrock (maximum 50m)	m	Pelletier et al.	-
soil_depth_statgso	soil depth (maximum 1.5m, layers marked as water and bedrock were excluded)	m	Miller and White (1998) - STATSGO*	-
soil_porosity	volumetric porosity (saturated volumetric water content estimated using a multiple linear regression based on sand and clay fraction for the layers marked as USDA soil texture class and a default value [0.9] for layers marked as organic material, layers marked as water, bedrock and "other" were excluded)	-	Miller and White (1998) - STATSGO*	Table 4 in Cosby et al. (1984), Lawrence and Slater (2008)
soil_conductivity	saturated hydraulic conductivity (estimated using a multiple linear regression based on sand and clay fraction for the layers marked as USDA soil texture class and a default value [36cm/hr] for layers marked as organic material, layers marked as water, bedrock and "other" were excluded)	cm/hr	Miller and White (1998) - STATSGO*	Table 4 in Cosby et al. (1984), Lawrence and Slater (2008)
max_water_content	maximum water content (combination of porosity and soil_depth_statgso, layers marked as water, bedrock and "other" were excluded)	m	Miller and White (1998) - STATSGO*	-
sand_frac	sand fraction (of the soil material smaller than 2 mm, layers marked as oragnic material, water, bedrock and "other" were excluded)	%	Miller and White (1998) - STATSGO*	-
silt_frac	silt fraction (of the soil material smaller than 2 mm, layers marked as oragnic material, water, bedrock and "other" were excluded)	%	Miller and White (1998) - STATSGO*	-
clay_frac	clay fraction (of the soil material smaller than 2 mm, layers marked as oragnic material, water, bedrock and "other" were excluded)	%	Miller and White (1998) - STATSGO*	-
water_frac	fraction of the top 1.5m marked as water (class 14)	%	Miller and White (1998) - STATSGO*	-
organic_frac	fraction of soil_depth_statgso marked as organic material (class 13)	%	Miller and White (1998) - STATSGO*	-
other_frac	fraction of soil_depth_statgso marked as other (class 16)	%	Miller and White (1998) - STATSGO*	-

5 **Table 6: Geological characteristics.**

Attribute	Description	Unit	Data source	References
geol_class_1st	most common geologic class in the catchment	-	GLiM	Hartmann and Moosdorf (2012)
geol_class_1st_frac	fraction of the catchment area associated with its most common geologic class	-	GLiM	Hartmann and Moosdorf (2012)
geol_class_2nd	2nd most common geologic class in the catchment	-	GLiM	Hartmann and Moosdorf (2012)
geol_class_2nd_frac	fraction of the catchment area associated with its 2nd most common geologic class	-	GLiM	Hartmann and Moosdorf (2012)
carb_rocks_frac	fraction of the catchment area characterized as "Carbonate sedimentary rocks"	-	GLiM	Hartmann and Moosdorf (2012)
geol_porosity	subsurface porosity	-	GLHYMPS	Gleeson et al. (2014)
geol_permeability	subsurface permeability (log10)	m^2	GLHYMPS	Gleeson et al. (2014)

Table 7: Main differences and similarities between MOPEX and CAMELS

	MOPEX	CAMELS
Number of catchments and spatial distribution	438 catchments, principally from the Eastern half of the CONUS and with an under-representation of the Rocky Mountains	671 catchments, with a relatively even distribution over the CONUS
Number of catchments in common to MOPEX and CAMELS	52	
Period covered by the hydro-meteorological time series	1948 - 2003	1980 - 2015
Minimal anthropogenic influence and long streamflow records assessed by:	Wallis et al. (1991) and the hydro-climatic data network (HCDN, Slack and Landwehr, 1992)	Updated version of the HCDN (HCDN-2009, Lins, 2012)
Raingauge density used as catchment selection criterion	Yes	No
Meteorological data	Point observations from National Climate Data Center daily Cooperative Observer Network (COOP) and SNOTEL stations. Long-term precipitation averages computed using 1961-1990 PRISM data	Gridded data from Daymet (Thornton et al., 2012), Maurer (Maurer et al., 2002) and NLDAS (Xia et al., 2012)
Streamflow data	USGS streamflow measurements	
Soil data	STATSGO (Miller and White, 1998)	STATSGO (Miller and White, 1998) and Pelletier et al. (2016)
Vegetation data	North America Land Data Assimilation (NLDAS)	MODIS imagery
Geology data	Not available	GLiM (Hartmann and Moosdorf, 2012) and GLHYMPS (Glesson and et al., 2014)
Reference papers	Duan et al. (2006) and Schaake et al. (2006)	Newman et al. (2015) and this paper