Title
Quantifying the effects of land use change on hydrology in Brazil

Author
Levy, Morgan Campbell

Publication Date
2016

Peer reviewed|Thesis/dissertation

Quantifying the effects of land use change on hydrology in Brazil

by

Morgan Campbell Levy

A dissertation submitted in partial satisfaction of the

requirements for the degree of

Doctor of Philosophy

in

Energy and Resources

in the

Graduate Division

of the

University of California, Berkeley

Committee in charge:

Professor Sally E. Thompson, Chair
Professor Laurel G. Larsen
Professor Maximilian Auffhammer

Fall 2016

Quantifying the effects of land use change on hydrology in Brazil

Abstract

Quantifying the effects of land use change on hydrology in Brazil

by

Morgan Campbell Levy

Doctor of Philosophy in Energy and Resources

University of California, Berkeley

Professor Sally E. Thompson, Chair

Quantifying the effect of changes in the Earths surface on regional water cycles is essential for water security. Hydrologists have traditionally used manipulative experiments in individual or paired river basins, or model-based analyses, to identify water cycle responses to land use or land cover change. Limitations in these approaches leave important gaps in understanding: (i) existing studies are not representative of biomes worldwide, and there is a deficit especially for tropical regions; (ii) individual site-specific experiments generally do not provide a representative sample of regional river systems in a way that can inform policy; and (iii) regional-scale analyses are largely based on simulation models, and are therefore limited by parameterizing assumptions and calibration uncertainty.

Larger-sample, empirical analyses are needed for more accurate modeling and policy-relevant understanding and these analyses must be supported by regional data. The Amazon-Cerrado (tropical savanna) transition region in Brazil is a global agricultural and biodiversity center, where regional climate and hydrology are projected to have strong sensitivities to land cover change. Dramatic land cover change has and continues to occur in this region, and its effects on streamflow (as an overall indicator of water cycle function) are not empirically understood at regional scales in part due to data uncertainties. Therefore, analyses of hydroclimate in the Amazon-Cerrado region of Brazil exhibit many of the challenges of interest to evaluation of regional hydrological change, as well as a suite of important applications.

This thesis explores data uncertainties implicit in the study of hydrological systems, and the hydrological effects of anthropogenic land use change. Specifically, this thesis: describes a novel data collection effort supporting empirical analysis at multiple-basin scales in Amazon-Cerrado Brazil; evaluates rainfall data uncertainty embedded in the study of the hydroclimate; and measures the effect of agricultural-driven deforestation on regional streamflow. This work required the harmonization of multiple in-situ and remotely-sensed (e.g. satellite-derived land use and climate) data products, and novel application of empirical statistical analysis methods developed in other fields (i.e. public health and economics)

to establish causality in complex observational data settings. At the time of writing, this research is the first application of these methods to a geoscience inquiry.

Firstly, this research contributes a novel hydrological dataset, including processing and quality control of more than 1,000 rain gauges, over 300 streamflow gauges, and associated GIS data (rain and streamflow gauge locations, river basin delineations, and large reservoir locations and drainage area delineations) across eight states in Brazil. Secondly, the research demonstrates that the magnitude of uncertainty from rainfall "data selection uncertainty", or uncertainty across multiple in-situ and remotely-sensed (satellite) rainfall data products, is comparable to estimated bias in global climate model projections, and provides practical recommendations for addressing this problem. Third, the research provides a series of regression analysis-based quantifications of the causal effects of deforestation on stream flow, which show that (a) streamflow, and especially ecologically-important dry-season low flow, has significantly increased across Amazon-Cerrado Brazil, and (b) that annual average increases in streamflow due to land cover change (agricultural development and corresponding forest loss) accounts for nearly half of total streamflow increases in the region over the past half century.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Background

Across Brazil's arc of deforestation, the northward-moving agricultural frontier located along the edges of the Amazon and Cerrado (tropical savanna) biomes, large-scale deforestation for pastureland and cropland began as early as the 1960s, peaked in the 1980s and 1990s, continued but slowed in the 2000s, and intensified yet again after 2012 [1]. Replacement of natural vegetation, including forest and tropical savanna woodlands, with pasture and cropland reduces evapotranspiration (ET) [2, 3], which has the primary, near-term effect of increasing streamflow [4, 5, 6, 7, 8, 9]. Transition of deep-rooted forest to shallower-rooted grassland and (rainfed) agricultural vegetation can alter not only the magnitude of ET, but also its seasonal pattern; deeper-rooted plants can continue to access soil water during dry periods whereas shallower-rooted plants cannot [10, 11, 12]. Thus, deforestation has the capacity to increase flow during dryer (low flow) periods in particular. Additional effects of land use change on streamflow include changes to soil hydraulic properties, such as reduced infiltration in pastureland or increased infiltration in cropland [6, 13]; increased sediment flux, particularly in large river basins [14]; and climate feedbacks - when reduced ET is coupled with climate change, rainfall may either increase or decrease, and it's seasonal pattern may change [15, 16, 17, 18, 19, 20, 21].

The effects of deforestation on streamflow are known to vary across basins and are scale-dependent [4, 14, 16], and existing studies hypothesize or suggest regional-scale effects based on individual basin findings. There remains limited empirical understanding of the direct effects of large-scale deforestation on streamflow at regional, multiple-basin scales in Brazil. Yet, environmental, agricultural, and energy policy efforts require understanding. River runoff volumes and timing affect hydroelectric power generation, navigation, the availability of freshwater for human consumption and agricultural production, and stream temperature and water quality that are critical to the health of uniquely biodiverse aquatic and terrestrial ecosystems [22, 23].

Brazil ranks third in the world in renewable electricity production, and is the world's

second largest hydropower producer [24]. Small hydropower capacity is slated for growth in developing rural areas [24], and the majority of the region's 243 planned large reservoir facilities will be located in the high-deforestation region of the southern Amazon [25]. Climate change and indirect climate feedbacks threaten to reduce streamflow and hydroelectric production - especially for drought-sensitive small hydropower facilities - thereby threatening energy security [26, 27]. Changes in streamflow could also potentially affect navigability of the 13,000 km of inland waterways currently transporting 45 million tons/year in agricultural and industrial goods [28]. Deforestation can increase sediment transport, particularly in large river basins [14], and resulting infrastructure and ecosystems effects are unknown. In this region, aquatic and terrestrial life is sensitive to water quality, quantity, and velocity, all of which are governed by streamflow volumes and timing. The Amazon is home to 2,320 known fish species, 64% of which are endemic [29]. Fish, aquatic invertebrates, and migratory birds rely on seasonal flood cycles; and large migratory fish are sensitive to connectivity of headwater breeding grounds [25]. Changes to the seasonal connectivity of rivers and floodplains due to streamflow change could disrupt highly specialized ecosystems [30]. Lastly, Brazil's economic future depends on continued production and export of soybean, maize, and beef - all of which are reliant on the seasonal climate of the southern Amazon and Cerrado, and continued agricultural growth and intensification will likely involve irrigation development [1]. Long-term reduction in rainfall and surface water availability due to indirect effects of deforestation would challenge Brazil's agricultural productivity [18].

Chapters

Curated Rain and Flow Data for the Brazilian Rainforest-Savanna Transition Zone

Description of data acquisition and processing are a necessary 'data-scientific' component of empirical research on hydroclimate and land use changes in Amazon-Cerrado Brazil. Thus, **Chapter 2** describes a custom data package entitled "Curated rain and flow data for the Brazilian rainforest-savanna transition zone", which provides a novel, curated set of long-term, historical daily rainfall and streamflow data from a large region spanning the southern Amazonian rainforest to Cerrado (tropical savanna) biomes of Brazil. The curated data set was derived from rainfall, streamflow, and associated geographic information systems (GIS) data obtained primarily from the Brazilian government water management agency, and also from the government electricity regulatory agency: Agência Nacional de Águas (ANA) and Agência Nacional de Energia Elétrica (ANEEL), respectively. The data package has been made available online via Figshare [31], and an analysis-ready subset of the full data are made discoverable through the Consortium of Universities for the Advancement of Hydrologic Science (CUAHSI) Hydroshare platform [32]. Curation of quality-controlled climate and hydrological data is an important but often overlooked 'data scientific' contribution to environmental research [33]. Because data scientific work can be a chief area of effort in

research on environmental change, explicit presentation of data acquisition and processing methods acknowledges the challenges posed by the structure, size, messiness, and complexity of data [34, 35], and the unique knowledge set required to address data challenges - especially with respect to data from rural developing regions of the world.

Addressing rainfall data selection uncertainty using connections between rainfall and streamflow

Studies of the hydroclimate at regional scales rely on spatial rainfall data products, derived from remotely sensed (RS) and in-situ (IS, rain gauge) observations. Because regional rainfall fields cannot be directly measured, these data products contain artifacts, which cause biases in their representation of the rainfall process. These biases pose a potential source of uncertainty in environmental analyses, attributable to the choices made by data-users in selecting a regional representation of rainfall. **Chapter 3** uses the rainforest-savanna transition region in Brazil as a case study and show differences in the statistics describing rainfall across nine remotely-sensed (RS) and interpolated in-situ (IS) daily rainfall datasets covering the period of 1998-2013. These differences propagated into illustrative analyses exploring temporal trends in monthly rainfall, and the computation of descriptive hydroclimate indices. The magnitude of the differences was large enough to potentially bias interpretation of environmental behavior. For instance, differences between rainfall datasets were comparable to estimated bias in global climate model projections, and rainfall trends from different datasets were inconsistent at the scale of river basins. To address this uncertainty, we evaluated the correspondence of different rainfall datasets with streamflow from 89 river basins to guide rainfall data selection. We demonstrate that direct empirical comparisons between rainfall and streamflow provide a scalable alternative to modeling for evaluating rainfall dataset performance across multiple areal (river basin) units. These results highlight the need for users of rainfall datasets to quantify this "data selection uncertainty" problem, and either justify data choices for hydroclimatological analyses, or report the uncertainty in derived results.

Land use change increases streamflow across the arc of deforestation in Brazil

Nearly half of recent decades' global forest loss occurred in the Amazon and Cerrado (tropical savanna) biomes of Brazil. Despite individual basin experimental and model-based analyses, a regional and empirical understanding of the direct effect of deforestation on streamflow is lacking. Streamflow is a key indicator of water cycle function, and streamflow regimes support globally-important ecosystems and industry. **Chapter 4** frames the case of land use change in Brazil as a natural experiment, and produces an estimate of the direct causal effect of deforestation on sreamflow within an observational data setting. Using a difference-in-differences (DID) regression modeling approach, we find that deforestation is responsible

for significant increases in streamflow across the Amazon-Cerrado region, specifically during periods of low flow. 2000-2013 flow increases in basins with high levels of deforestation were between 5 and 11 percentage points higher than in minimally deforested basins. A mixed effects regression model that estimates the relationship between annual rates of flow and different levels of agricultural development and forest cover indicates that between 1950-2012, the Amazon-Cerrado region experienced an increase of 0.93 mm/year, or 2.94 km^3/year, due to land cover change, which accounts for 44% of total regional flow rate increases (6.72 km^3/year), with remaining increases due to climate.

Chapter 2

Curated rain and flow data for the Brazilian rainforest-savanna transition zone

2.1 Introduction

The "Curated rain and flow data for the Brazilian rainforest-savanna transition zone" package (hereafter 'data package') provides a curated set of long-term, historical (intermittent, 1926 - 2013) daily rainfall and streamflow data from a large region spanning the southern Amazonian rainforest to the Cerrado (tropical savanna) biomes of Brazil (see Figure 2.1). The curated data set was derived from free, publicly-available rainfall, streamflow, and associated geographic information systems (GIS) data obtained primarily from the Brazilian government water management agency, and also from the government electricity regulatory agency: Agência Nacional de Águas (ANA) and Agência Nacional de Energia Elétrica (ANEEL), respectively. Details of data formatting and quality assurance/quality control (QA/QC) are provided here. Programmatic scripts and data files referenced in this document are included in the associated data package, which has been made publicly available online via Figshare [31]; an analysis-ready subset of the full data are made discoverable through the Consortium of Universities for the Advancement of Hydrologic Science (CUAHSI) Hydroshare platform [32].

This manuscript makes reference to files contained within the data package, and complete information on data package files and data structures are included in a "ReadMe" file located at the Figshare instance of the data package. This manuscript is intended to fully document the data acquisition and processing effort as a transparent 'data-scientific' component [33] of the larger research on hydroclimate and land use changes in Amazon-Cerrado Brazil.

2.2 Data Acquisition

Time Series Data

I downloaded in-situ daily rainfall intensities or depths [mm/day] (hereafter 'rain') and daily average streamflow rates [m^3/sec] (hereafter 'flow') from the ANA historical data web portal [36] (CSV). I also downloaded corresponding geographic information system (GIS) data including rain and flow gauge site locations (KML) from this same source (see Section 2.2). Data were selected and downloaded by state, for the following states: Acre (AC), Amazonas (AM), Gois (GO), Mato Grosso (MT), Mato Grosso do Sul (MS), Para (PA), Rondnia (RO), and Tocantins (TO). I obtained active as well as historical (inactive) rain and flow data. Selected rain sites include all sites in these states; selected flow sites include only sites located within a custom study region. The custom study region cuts the northern states (AM and PA) at the Amazon River, and cuts the southern-most state MS along a line defined by the bounding-box of an original four-state (MT, RO, GO, TO) region of interest. GIS data used to delineate the custom study region include Brazilian river courses and political boundaries, which I downloaded from the ANA Geospatial Metadata Portal [37]. The custom study region is shown in Figure 2.1. Data for the states of MT, RO, GO, and TO were downloaded on 12/5/2014; data for the states of PA, AM, and MS were downloaded on 3/25/2015; data for AC was downloaded on 3/27/2015.

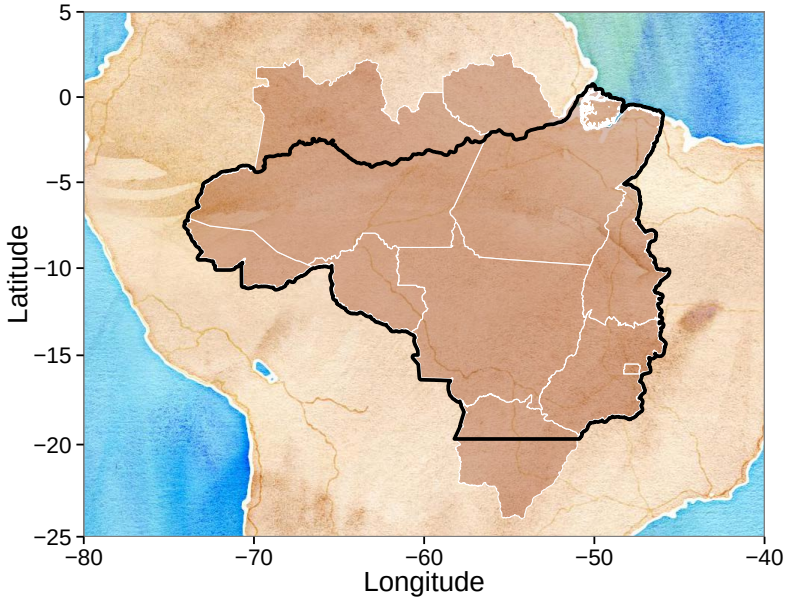

Figure 2.1: Study region of Brazil

Outline of the custom study region (black line) in Brazil, overlapping the eight states (white lines, brown fill) within which rain and flow data were downloaded. Cleaned (analysis-ready) rain data is available in the full eight-state (brown) study region. Cleaned (analysis-ready) flow data is available in the subest (black line) study region. Data source: ANA Geospatial Metadata Portal [37].

I downloaded time series CSVs by state using a custom python script, which downloads a separate CSV file for each site in a selected group (e.g. in a selected state) and for an individual selected data type (e.g. rain or flow). Each site CSV covers a unique date range determined by data availability at each site. The earliest dates of record are in the 1920s, however most sites began reporting in the 1980s and some start as late as the early 2000s; at the time data were downloaded, the most recent data available were reported in late 2014. Records included in the package date no later than 12/31/2013.

Source Data Details

The information in this section describes data collection and pre-processing *carried out by The Agência Nacional de Águas (ANA).*

The Agência Nacional de Águas (ANA) measures total accumulated rainfall daily in conventional bucket-type gauges, and checks raw data for observer errors and consistency [38, 39, 40]. Consistency analysis of rainfall data starts with the determination of homogenous rainfall regions [41, 42, 43]. Within each homogeneous region, primary rainfall gauges are defined based on their location, length of time series, length of period of interruption, and record of changes in location or measurement equipment. Secondary gauges are used to

verify observation errors in data from the major gauges, and are defined based on correlation (higher than 0.8) and distance (less than 200 km) with respect to the major gauges. Primary and secondary gauges must have coincident periods of observation and be located at similar altitudes. Four methods are used to compare and verify inconstancies in monthly and annual rainfall among gauges: regional weighting [44], linear regression [44], double mass curves [45], and regional vector [46, 47]. When inconsistencies in monthly and annual rainfall are clearly identified, daily values are corrected accordingly; otherwise, values are flagged as missing for the entire inconsistent period.

To obtain streamflow data, the ANA measures river water stage twice daily, and checks raw data for observer errors. Field teams take river flow measurements once in every four-month period, and those raw data are checked for consistency between cross section profiles, water level depths, wetted areas, and measured river velocities. Streamflow raw data is checked for consistency [48, 49, 50]. Procedures consist of visual assessment of consistency of river flows, specific river flows, relationships between specific river flows and drainage area, and flow duration curves - at daily, monthly and annual time scales, and river stage-flow curves. Consistency of the relationship between raw river stage and flow measurements is checked using dispersion plots. Using only consistency-checked data, river stage curves are constructed [50] and validated by inspection of differences between observed and calculated river flows plotted against river levels and time (must be below 20%). Extrapolation of river stage curves uses six different methods, according to the characteristics of river channels [51]: logarithmic, area versus water velocity, Stevens (using Chezys equation), Stevens (using Mannings equation), Manning, and river slope-conveyance. River flows computed using validated river stage curves are finally submitted to visual inspections, focused on checking the consistency of hydrographs of streamflow gauges located on the same river. Such visual checks include comparisons of daily and monthly streamflow and specific flow time series for different gauges, relationships between specific flows and drainage areas, verification of monthly incremental river flows (differences between downstream and upstream monthly flows should be positive), and daily and monthly flow duration curves for different gauges.

Raw and consistency-checked daily rainfall and streamflow are flagged in the publicly available data from the ANA historical data web portal [36]. The curated data described in the data package that accompanies this dissertation includes raw rainfall data, and a combination of raw (constituting 10% of final quality-controlled data) and consistency-checked (constituting 90% of final quality controlled data) streamflow data in order to obtain date ranges beyond those available for consistency-checked data alone (2006). Therefore, the author of this dissertation performed additional custom QA/QC - see Section 2.3.

Spatial Data

All spatial data in the data package are in geographic coordinates (latitude/longitude) on the World Geodetic System of 1984 (WGS84) datum. Note that as one move towards the Amazon, gauged locations for both rainfall and streamflow thin, and there is an association

of the location of gauge sites with road and/or river networks, due presumably to access issues.

Rain and Flow Site Locations

The ANA web portal also provides GIS location data for all rain and flow site data. These are available as KML files that can be downloaded along with time series data CSVs. I downloaded rain and flow site locations by state, along with the time series data (on the same dates). GIS files of all rain and flow gauge locations are included in the data package. Rain and flow location metadata attributes include site number (corresponding to the rain and flow site numbers in the time series data - see Section 2.3), state, and elevation in meters above sea level extracted from the NASA Shuttle Radar Topography Mission (SRTM) v.4 digital elevation model (DEM) [52].

Reservoir Locations

The locations of operational and under-construction hydroelectric sites (and therefore dam and/or reservoir facilities that could alter natural streamflow regimes) are available through ANEEL; the data I used are current as of 2013. These data were compiled and made easily-accessible at dams-info.org, and an original source version of these data were provided to us directly by Zachary Hurwitz at International Rivers, the organization that maintains the dams-info website. I attempted to download hydroelectric facility location and metadata directly from ANEEL's Sistema de Informações Georreferenciadas do Setor Elétrico (SIGEL) website [53] - the same source used by International Rivers, however the website's data export functions were non-functional in 2014 when these data were being collected. Hydroelectric facility location data alone is available from the SIGEL KMZ download website [54], however the metadata that is used to distinguish active and under-construction vs. planned sites, for example, was not.

I selected large hydroelectric facilities ("Usinas Hidreltricas" - UHE, which are categorized by ANEEL as facilities with a $>$ 30MW capacity) located in the study region of interest that were reported as under-construction or operational as of 2013. This means that the facility had the potential to impact natural streamflow, either through construction or operation, at some point prior to or during 2013 (additional research would be required to identify the actual start date of potential impacts to flow by individual hydroelectric facilities). Small hydroelectric facilities ("Pequenas Centrais Hidreltricas" - PCH, which are categorized by ANEEL as facilities with a $<$ 30MW capacity) were excluded from data processing and subsequent analyses, as they were assumed unlikely to substantially impact flow regimes (i.e. they were assumed to be small and/or run-of-river hydroelectric facilities). The data package includes a GIS file of active or under-construction large hydroelectric facility (reservoir) sites located in the custom study region. Reservoir location metadata attributes include a custom ID, state in which the site is located, two ID categories assigned by ANEEL (retained for

record-keeping purposes only), and elevation in meters above sea level extracted from the NASA Shuttle Radar Topography Mission (SRTM) v.4 digital elevation model (DEM) [52].

Watershed Boundaries

I downloaded GIS watershed boundary data, which I used to delineate flow site and reservoir site drainage areas, from the ANA Geospatial Metadata Portal [37]. Specifically, these data are a collection of high-level (detailed) river drainage areas (watersheds, basins or catchments) coded according to the Otto Pfafstetter stream/basin coding system. At any point along a river (such as at the location of a flow site or reservoir), these coded basins can be aggregated according to the Otto Pfafstetter codes to construct a complete drainage area for the selected point. This coding system, and the ANA's process for constructing the Otto Pfafstetter-coded basin GIS using a DEM, is documented in an agency manual ("Manual de Construo da Base Hidrogrfica Ottocodificada da ANA") available at the ANA Geospatial Metadata Portal [37].

I aggregated these coded basins according to a custom algorithm, and constructed drainage areas for both the flow sites and reservoir locations. This is described in more detail in Section 2.3. These data were downloaded on June 23, 2014. (The format in which these data are provided online has since changed and/or been updated by the ANA; the script used to import and format those data for our analysis is specific to the version I downloaded in 2014. Therefore, this script is not provided along with the data package, but the data I used and the script is available upon request.) The raw data are not included in the data package because they can be obtained directly from the ANA. However, the custom-delineated drainage area (basin) boundaries for the flow and reservoir sites are included in the data package - see Section 2.3.

2.3 Data Formatting and QA/QC

Rain and Flow Time Series Formatting

Using the Comprehensive R Archive Network (CRAN) [55] programming environment (Version 3) on both Apple and Windows operating systems, I reformatted and date-regularized CSV time series data using custom functions (code is included in the data package). These custom functions make use of quality codes (qc) assigned by the ANA. Raw ANA data is assigned one of two qc values: 1 for raw data and 2 for processed data. These qc codes are treated differently for rain and flow in the custom functions named above due to differences in how data from each quality code is made available on the ANA website. For rain data, both qc code values 1 and 2 span similar date ranges. However, for flow, qc=1 tends to contain latest/recent years (only), and qc=2 the historical record, sometimes including recent years, but sometimes not. Thus, the quality codes are handled differently.

For rain, I compared a selection of sites for which qc date ranges overlapped. Results showed that qc=2 data are nearly identical to qc=1, except for cases where qc=2 is missing

dates that qc=1 contains; qc=2 data occur over a much narrower date range (e.g. start later, and end earlier; qc=2 only goes up to 2006). Following this check, I use qc=1 data, and complete further manual quality checks. For flow, a mix of qc=1 and qc=2 are used; qc=2 is prioritized (used as the main data source); qc=1 data is used if no qc=2 data is available; where qc=1 data can fill in gaps in qc=2, qc=1 data is merged with qc=2 data.

This combination of data with different qc codes was justified using visual inspection of individual site data by quality code type, as well as by checking summary statistics across all data. Summary statistics (min, max, mean, median) were computed for state-level qc=2 data, and compared to state-level qc=1 data, and were shown to be consistent. Summary statistics for state-level data matched rain and flow statistics reported in the literature for the study region [56, 57].

reformatted, date-regularized rain and flow data are included in the data package; these data remain 'raw' as the only processing that has been done is merging of data with different qc codes as described above, reformatting of the original CSVs to date-regularized, long format, and aggregation of all sites into a single data object for each type (rain and flow). A separate 'ReadMe' file included in the data package provides information on data structure. There are a total of 1,364 rain sites and 744 flow sites.

Rain and Flow Time Series Data QA/QC

Rain and flow data are cleaned using different approaches. For each data set, I created quality control (QC) notes to record data quality information, and to subsequently guide data cleaning. (The abbreviation 'QC' used here for custom quality control information is different from the 'qc' codes used by the ANA to indicate raw vs. consistency-checked rain and flow data.) Notes are included in the data package. In the QC notes, files are flagged if they should ultimately *not* be used in any analysis due to quality concerns. Entries in the QC notes also describe characteristics of the data noted during both automated and manual checks of the data, and upon which filtering of the data can be done. A more detailed discussion of the QA/QC process for each data type is provided below.

I took a conservative or minimal-intervention approach to cleaning these data. This means that if there was no clear reason to suspect poor quality data, site data were not altered. Where there was clear reason to suspect poor quality data, the only modification made to those data was the setting of poor quality data values to "NA". I made no other value adjustments or interpolations. To identify poor quality data, I looked at time series visualization of the data, histograms, QQplots and Kolmogorov-Smirnov tests (comparisons between distributions of nearby site data and/or between two time periods in the same site's data). Other than automated tests run to identify sites with outlier/extreme values, the identification of poor-quality sites was visually-based. Suspect data included data that had time-series inconsistencies (e.g. noticeably different patterns over different periods of time, or relative to nearby sites), extreme (high) values, and blocks or duplicated patterns indicating manually-altered entries or bad gauges. A double mass curve (DMC) analysis [45] was used to confirm consistency of rain and flow sites relative to neighboring sties, and to identify

sites that may be excluded from future analyses depending on the context in which they are to be used - see Sections 2.3 and 2.3 below.

Rain

This section describes the QA/QC process for rain data, which results in the creation of the quality controlled time series data. I checked the raw rain data for outliers along several dimensions: high and low values, means, medians, and missingness. Sites with outlier values were noted as requiring subsequent cleaning. Where outlier values indicated a more systematic problem with the site (e.g. all values were outside a reasonable range, or the site showed highly irregular seasonal patterns), the site was flagged so as to be excluded from any analysis. I found no values less than zero, and the only universal adjustment I carried out during the cleaning process was to assign all values higher than 250mm a "NA" value because in all cases, values greater than 250mm appeared to be measurement and/or reporting error, not extreme events. This was evident by observing rainfall events over a site's full time series, and especially the values adjacent to the extremes: extreme values assumed to be errors tended to have adjacent missing values, very low or zero values instead of preceding storm events, and/or occurred in the dry season with no preceding or subsequent rainfall. Thus, the classification of an erroneous extreme is somewhat qualitative, however it is justified by context knowledge of regional rainfall cycles and extremes.

I visually inspected the time series of all sites, even those that did not have outstanding values according to the statistical summaries. Where a (subjective, visually-determined) baseline signature of rainfall at a given site (or from nearby sites) was evident - usually requiring ≥ 3 years at a site, and segments of a site's time series were inconsistent relative to that baseline, then inconsistent entries were set to "NA". In the case of an entire site's time series being inconsistent, the site was flagged, and an accompanying descriptive note was included in the rain QC notes.

The only other cleaning, besides replacement of bad values with "NA" and flagging of bad sites, was removal and/or combination of sites with duplicate coordinates: sites at the same location. For sites with duplicate coordinates that had non-overlapping date ranges, I kept both sites as-is. For co-located sites with overlapping date ranges, I inspected time series for consistency, selected a primary site (the site with the longest duration or highest quality data), and I filled in missing entries from that site with data from the other duplicate site. I then set the duplicate sites' complimentary data to "NA" and/or flagged the entire site for removal if there was complete overlap. If a duplicate site showed inconsistent or poor quality data compared to its primary co-located site, then I flagged the duplicate site. Thus, no duplicate sites have overlapping time ranges in the final cleaned data set. When entire years of data were set to "NA", generally the period was set between July 1st of each year (a dry season date occurring well before the onset of wet season transitional rainfall), unless existing breaks in data could be used, or other time ranges were appropriate to the data. Sites with durations less than a year or sites containing too many gaps to provide a baseline

signature were flagged. All sites that were manually cleaned are noted in the QC notes, and I also trim missing values at the start or end of a time series in the quality controlled data..

After data cleaning, I used a double mass curve (DMC) analysis to confirm general consistency of rain site data relative to neighboring site data. For each good quality (unflagged) site, I identified other rain sites within a 100 km radius to use as a reference comparison group. The 100 km radius was used based on an analysis of rain correlation distances across the entire study region: I fit a variogram model to a random sample of 1,500 individual days between 1980 - 2013 for all rain sites active on each day, and estimated the mean variogram range (maximum correlation distance) to be 100km. Not all sites had other sites within this range and with adequate data due to missingness. Thus, it was not possible to perform a DMC on all sites. Site data in years with > 10 consecutive days of missing values were excluded from the analysis, for both the site and its reference group. In years with acceptable missingness (< 10 consecutive days of missing values), missing values that remained were filled using a 5-day moving average value (for the DMC analysis only, not in the quality controlled data). The number of reference group sites varied by year for each site due to missingness. Out of a total of 1,171 good quality sites, it was possible to do a DMC analysis on 1,058 sites (meaning there was at least one other site within 100km that had at least one year of data with < 10 consecutive days of missing values). Sites for which a DMC analysis was performed are noted in the QC notes.

DMCs were created by individual years due to uneven reporting (missingness) across sites. To evaluate the consistency of a site over all years relative to its reference group, I fit a linear regression (with an intercept forced to the origin) of the cumulative sum of the site's daily rainfall on the mean cumulative sum of the reference group's daily rainfall. This was done separately for each year available. Not all years in a site's record could be evaluated due to missingness in either the site itself, or its reference group sites. I calculated summary statistics of a site's DMC fit across all years for which I was able to construct a DMC. These statistics included the mean and median of residuals, the mean standard deviation of residuals, and the mean R^2 fit of the linear regression across all individual-year DMC fits combined.

The R^2 interquartile range (IQR) for all sites was between .993 and .996, meaning that outliers with respect to this statistic were not especially meaningful; anything with an R^2 less than 0.987 was an outlier, and the lowest R^2 was 0.84. The most meaningful statistic is the mean of residuals across all years' DMCs for a site, which is centered at zero for most sites. Sites with mean residuals that qualify as outliers may have systematic bias in their rainfall data relative to their reference groups. There were 92 outlier sites - those with residual mean values exceeding a value equal to 1.5 times the IQR upper and lower limit. I found that outlier sites were sites for which the number of years for which DMCs were calculated was low relative to the average across all sites (the distribution of the count of years used to calculate DMCs for the outlier sites was shifted to the left of the distribution for all sites). Thus, the outlier nature of these sites may be attributable to limited data, and not necessarily to a real lack of consistency. 67% of the sites identified as outliers according to their residual standard deviation were also identified as outliers according to their residual mean. Sites

that were identified as outliers according to either the mean or standard deviation of the residuals were noted in the QCnotes. A total of 103 sites were noted as outliers according to the DMC analysis (99 sites according to the residual mean, and an additional 11 unique sites according to the standard deviation of the residuals).

Despite being statistical outliers, these sites may not necessarily need to be removed from analysis due to the fact that their outlier status may be a result of limited data, rather than the quality of the data. The DMC analysis in general confirms site consistency more than it does indicate poor quality sites. Outlier site flags remain set to zero (indicating good quality). It is up to the user to decide whether or not to exclude sites based on the DMC results.

Traditionally, DMC analyses are used for identification of inconsistent gauge data where local, context information is often available for individual sites, and where detailed adjustments can be made to individual sites using results from a DMC analysis. In the context of this relatively large data set, the use of a detailed DMC analysis was not feasible, and therefore the summary statistical approach was appropriate. The testing of significance of DMC residuals generally relies on identification of change-point years (based on visual inspection of DMCs and context knowledge); this was not carried out. Therefore results were used to summarize DMC fits to confirm that on average, these sites are consistent with their neighboring sites. Additionally, the sites were compared to other reference sites in a large (100 km radius) region; correction would not necessarily be desired due to uncertainty in the reasons for site inconsistency across this great of a distance.

Lastly, I checked for inconsistencies in daily rainfall values by weekday, and specifically for systematically different reporting on weekend days (Saturday and Sunday). I checked the proportion of missing values on each day of the week; standard deviation of daily rainfall (as a fraction of mean daily rainfall) on each day of the week; and proportion of zero values on each day of the week for all un-flagged sites. There were no apparent inconsistencies for the standard deviations or proportion of zero values on weekend days as compared to weekdays (or on any individual day compared to other days of the week). There were, however, six sites with a significantly greater number of missing values on weekends. These sites are noted in the QC notes, but not flagged because their weekday data is still usable. A greater number of missing values on weekends should not affect summary statistics of rainfall, or by-day interpolations, other than to reduce the number of observations available.

Final, quality-controlled daily rainfall data (for un-flagged sites only) is provided in the data package separate from the raw data, and are the data made discoverable through the CUAHSI platform. All notes on cleaning remain in rain QC notes, which can be used to filter rain data for analysis (based on flag values, at a minimum). Any unflagged rain site has been determined by the above-described QA/QC process to be ready for use in analysis. Of the original 1,364 (raw) rain sites, 1,171 are determined to be adequate quality for analysis, however use of other filters (e.g. duration of reporting, being an outlier according to the DMC analysis) may further limit the number of sites available for analysis. These sites have extremely varied reporting periods and missingness. Figure 2.2 shows all rain site locations in the study region.

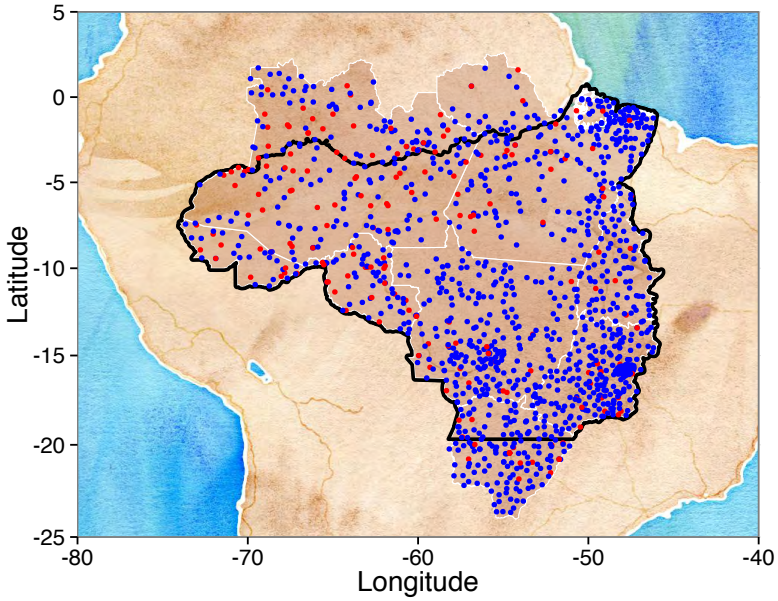

Figure 2.2: Rain site locations

Blue indicates good quality sites (unflagged) and red indicates poor quality sites (flagged). Data source: ANA historical data web portal [36].

Table 2.1 summarizes the information provided in the rain QC notes. These notes were created in the process of QA/QC, and many refer to issues that were identified in the original, raw data, and do not remain in the final, cleaned data (but are retained for record-keeping). All issues identified in the notes were addressed in the cleaning process and/or sites were flagged as poor quality when issues could not be addressed; only flagged sites have remaining problems. These notes are intended to be used alongside the final cleaned data for filtering of data for specific analyses.

Table 2.1: QA/QC notes for rain data

Note	Description
<X year	Sites with data over a period less than X number of years, e.g. '<1 year'.
>X% values over Y	A site with a percentage X of values exceeding some threshold value Y; this note was used as an indicator for further inspection and potential cleaning of a site.
cut/clean	Sites that required detailed checking (some notes contain specific date ranges or refer to specific ANA quality codes over which to check data), cleaning (setting of some data to 'NA'), and potential flagging; sites that also have 'clean_manual' = 1 were cleaned; sites that also have 'flag' = 1 were flagged.
DMC	A double-mass curve (DMC) analysis was performed on this site.
DMC outlier	A double-mass curve (DMC) analysis determined this site to be an outlier according to summary statistics across all sites.
duplicate of X	Sites that are located in the same location as other sites (at the same or different time periods), where X is the site ID number of the co-located site; modifications of this note might include information on how duplicate sites were handled, e.g. 'moved overlap to duplicate X', or 'NAs replaced with duplicates from X'.
max value >300 or max value >= 250	Sites with original (not-cleaned) data values >250 or >300 [mm/day]; all values >250 were set to 'NA' during cleaning based on summary statistical analyses and context knowledge (regional climate); if issues remain with the site, it will be flagged (flag = 1).
mean >= X, or mean (median) on rain days >= X	Sites with a mean or median value (over the full data record, or only for days in which there was non-zero rainfall) determined to be greater than or equal to a value X considered to be an outlier according to summary statistics; this note was used as an indicator for further inspection and potential cleaning of a site.
more NAs on weekends	Sites that report a greater number of missing values on weekend days (Saturday, Sunday) than on other days of the week.
NA values >X% of data	A site with a quantity of missing values that exceed a threshold percentage X of the data; this note was used as an indicator for further inspection and potential cleaning of a site.
potentially inconsistent or inconsistent	A site with data that appeared inconsistent due to real (climate) variation or measurement error - it was not possible to distinguish.
abnormal	A site deemed to be poor quality upon visual inspection; an indicator that the site was checked, was not able to be cleaned so as to be categorized as good quality, and was therefore flagged (flag = 1).
zero values >X% of data	A site with the percentage of values = 0 exceeding some threshold X; this note was used as an indicator for further inspection and potential cleaning of a site.
(other)	Other custom notes may be appended to the above common notes, or may be entered independently; e.g. 'drought 1990s', which is information relevant to the site and provided for context.

Flow

This section describes the QA/QC process for flow data. For flow data, quality controlled sites were limited to those within the custom study region area; this eliminated sites located in the northern parts of AM and PA (above the Amazon river) and MS to the south (below the latitude of other included state areas). The total number of sites contained in the study region is 626, however not all of these were investigated due to the presence of reservoirs,

described in more detail below. A total of 611 sites have QA/QC information provided in the flow QC notes (see data package).

I first identified which flow sites were located in or immediately downstream of a large reservoir facility. This was done initially in order to exclude sites that are directly impacted by reservoirs: flow measurements made in or just downstream of a reservoir do not provide a natural flow record. To identify sites located at reservoirs, I looked at satellite images and maps of flow gauge and reservoir locations on the ANA historical web portal from which the original data were obtained; the ANA web portal visualizations sources Google satellite imagery dated 2014. I also looked at the locations of operational and under-construction dam sites listed by ANEEL as of 2013 using the dam location data available from ANEEL directly, and also courtesy of International Rivers (see Section 2.2). Flow sites that are located immediately downstream of a reservoir or appear to be located in a reservoir area are noted in the flow QC notes. In combination with this analysis, drainage basins for each flow site and reservoir facility were identified using the watershed GIS data obtained from the ANA Geospatial Metadata Portal [37] (see Section 2.3). The areas draining to each flow site (its catchment area in square km), as well the percentages of those areas that are composed of any reservoir's drainage area, are noted in the flow QC notes.

For flow data, I made no universal adjustments; all adjustments were based on visual inspections of the time series data of each site. I also checked flow data for outliers along several dimensions similar to rain: high and low values, means, medians, and missingness. I used these criteria to find individual sites that required cleaning, or that needed to be flagged as poor quality (and not used in analysis). I looked at hydrographs (flow time series) at each flow site and noted where non-stationarities, uncharacteristic trends, or missingness were evident. I flagged data that appeared inconsistent or otherwise poor quality, and that could not be adequately trimmed or adjusted.

Adjustments made to data include replacement of bad values with "NA"", and removal and/or combination of duplicate sites, similar to the rain data. Duplicate flow sites were identified by checking site coordinates as well as overlap in drainage basins. For flow sites with non-overlapping date ranges, I kept both sites as-is. For sites with overlapping date ranges, I inspected time series for consistency, selected a primary site (the site with the longest duration or highest quality data), and I filled in missing entries from that site with data from the other duplicate site. I then set the duplicate sites' complimentary data to "NA" and/or flagged the entire site for removal if there was complete overlap. If a duplicate site showed inconsistent or poor quality data compared to its primary co-located site, then I flagged the duplicate site. Thus, no duplicate sites have overlapping time ranges in the final cleaned data set. Missing values at the start or end of a time series were trimmed.

After data cleaning, I used a double mass curve (DMC) analysis to confirm general consistency of flow site data relative to neighboring site data. I used the same process as described for the rain data (see Section 2.3), with some modifications to the process specific to flow data. Analysis-ready flow sites are those with good quality (unflagged) data, no reservoir onsite, with less than 10% of their total drainage area affected by a reservoir, and sites for which basin areas were available (meaning the basin had been delineated - see Section

2.3). Additionally, volumetric flows were converted to area-normalized depths (mm/day) for the DMC analysis.

For each of these selected sites, I identified other flow sites within a 100 km radius (again, based on rain correlation distances across the entire study region). It was not possible to perform a DMC on all sites. Site data in years with > 10 consecutive days of missing values were excluded from the analysis, for both the site and its reference group. In years with acceptable missingness (< 10 consecutive days of missing values), missing values that remained were filled using a 3-day moving average value (slightly smaller than the window used for rain due to increased missingness in the flow data). The number of reference group sites varied by year for each site due to missingness. It was possible to do a DMC analysis on 279 sites (meaning there was at least one other site within 100km that had at least one year of data with < 10 consecutive days of missing values). Sites for which a DMC analysis was performed are noted in the flow QC notes.

The R^2 interquartile range (IQR) for all flow site DMC fits was even tighter than for rainfall. An outlier site was any site with a mean R^2 less than 0.988, again meaning that outliers with respect to this statistic were not especially meaningful. As in the rain DMC analysis, the most meaningful statistic appeared to be the mean of residuals across all years' DMCs for a site, which is centered at zero for most sites. Sites with mean residuals that qualify as outliers may have systematic bias in their streamflow data relative to their reference groups. There were 85 outlier sites according to this statistic - those with residual mean values exceeding a value equal to 1.5 times the IQR upper and lower limit. 98% (all but one) of the sites identified as outliers according to their residual standard deviation were also identified as outliers according to their residual mean. Sites that were identified as outliers according to either the mean or standard deviation of the residuals were noted in the flow QC notes. A total of 86 sites were noted as outliers according to the DMC analysis (85 sites according to the residual mean, and an additional 1 unique site according to the standard deviation of the residuals).

Unlike with rainfall, outlier flow sites appeared no different than all sites in terms of the number of years over which DMC residuals were calculated. Nor did basin size appear correlated with categorization as an outlier site. Therefore, the qualification of flow sites as outliers may potentially have more meaning than for the rain sites (where it was determined outlier status was correlated with limited years over which to calculate a DMC.) Again, the DMC fits for flow were very good, meaning also that outliers are potentially an artifact of the large number of sites available for this analysis, rather than something with a physical interpretation. Furthermore, the inconsistency of some basins relative to others might in fact be a signature of interest. Outlier flow sites should be excluded based on the analysis at hand. If flow is used as a reference case (e.g. for establishing relationships between rain and flow over time), it may be more desirable to remove the outliers from the analysis. If, however, inconsistency in a flow record in response to basin disturbance (e.g. land use change) is a signature of interest, it may be more desirable to retain the outliers in an analysis because no other significant problems were identified for these sites.

Final, quality controlled daily flow data (for un-flagged sites only) is provided in the

data package separate from the raw data, and are the data made discoverable through the CUAHSI platform. All notes on cleaning remain in flow QC notes, which can be used to filter flow data for analysis (based on flags, at a minimum). Any unflagged flow site has been determined by the above-described QA/QC process to be ready for use in analysis. Figure 2.3 shows all flow site locations in the study region.

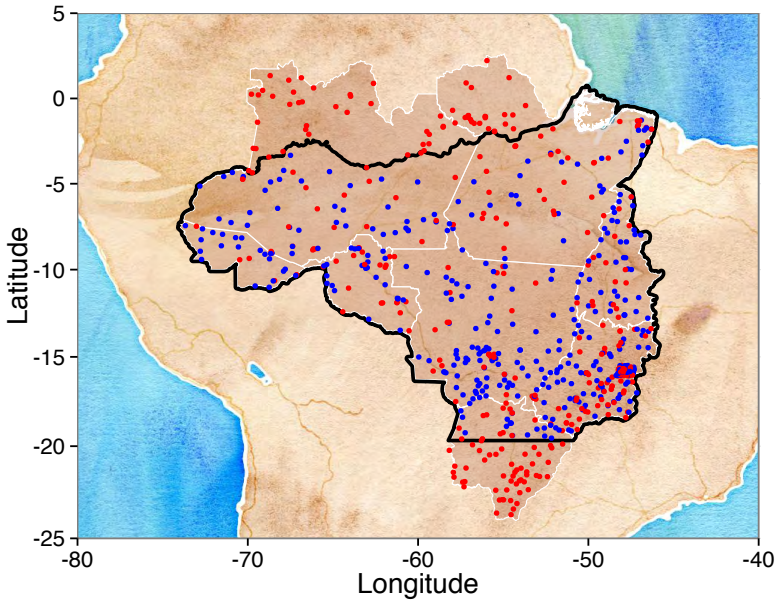

Figure 2.3: Flow site locations

Blue indicates good quality sites (unflagged) whose upstream drainage areas (basins) have been delineated; red indicates poor quality sites (flagged) or sites excluded due to other factors: presence of a reservoir, upstream impact of a reservoir, or a site located outside the custom study region. Data source: ANA historical data web portal [36].

Table 2.2 summarizes the information provided in the flow QC notes. These notes were created in the process of QA/QC, and many refer to issues that were identified in the original, raw data, and do not remain in the final, cleaned data (but are retained for record-keeping). All issues identified in the notes were addressed in the cleaning process and/or sites were flagged as poor quality when issues could not be addressed; only flagged sites have remaining problems. These notes are intended to be used alongside the final cleaned data for filtering of data for specific analyses.

Table 2.2: QA/QC notes for flow data

Note	Description
Bolivian or Peruvian basin	The site is located in a basin that extends into Bolivia or Peru.
cut/clean	Sites that required detailed checking, cleaning (setting of some data to 'NA'), and potential flagging; sites that also have 'clean_manual' = 1 were cleaned; sites that also have 'flag' = 1 were flagged.
decreasing	A site with data that has a decreasing or downward trend over time.
DMC	A double-mass curve (DMC) analysis was performed on this site.
DMC outlier	A double-mass curve (DMC) analysis determined this site to be an outlier according to summary statistics across all sites.
downstream from/of, or in, a reservoir or ROR	A site located downstream of or in the immediate vicinity of a reservoir or run-of-river (ROR) hydropower facility; modifications of this note include '(far) downstream of a reservoir', meaning assumed non-direct impact of the reservoir which is captured by 'reservoir_impact_p'; several sites' data precede construction of the co-located reservoir facility, and are noted e.g. 'data prior to reservoir construction'.
duplicate of X	Sites that are located in the same location as other sites (at the same or different time periods), where X is the site ID number of the co-located site; modifications of this note might include information on how duplicate sites were handled.
errors	Errors were noted in a site's data (that were likely addressed through cleaning in the final data set; if not, then the site will have flag = 1), e.g. 'min value errors' or 'max value errors'.
gap(s)	Indicates the presence of gaps (missing values) at any point in the time series; modifications of this note include 'small gaps', 'large gaps', and 'gap at X' where X is a year or date range.
good example site	Sites that had a consistent long-term record, and/or were located in specific region of interest, and were noted as potentially being useful in the capacity of a case study or example.
increasing	A site with data that has an increasing or upward trend over time.
low flow	Sites with relatively low flow volumes (small, potentially-headwater streams).
ok	A site deemed to be good quality upon visual inspection; an indicator that the site was checked, and was not flagged for further detailed inspection; this note is sometimes accompanied by specific reference to elements of a site's data, e.g. 'max values ok'.
potentially filtered	Sites that appeared upon visual inspection to potentially contain synthetic or modified data; sites that also have 'clean_manual' = 1 were cleaned; sites that also have 'flag' = 1 were flagged.
(potentially) inconsistent	A site with data that appeared inconsistent due to real (natural or basin-changed induced) variation or measurement error - it was not possible to distinguish.
small date range	A site with data over a relatively small date range; usually accompanied by a note on the number of years (e.g. '<10 years', '<5 years', or '<2 years').
abnormal	A site deemed to be poor quality upon visual (hydrograph) inspection; an indicator that the site was checked, was not able to be cleaned so as to be categorized as good quality, and was therefore flagged (flag = 1).
(other)	Other custom notes may be appended to the above common notes, or may be entered independently; these are mostly self-explanatory.

Drainage Area Formatting

I aggregate the ANA Otto Pfafstetter-coded basins (see Section 2.2) into flow- and reservoir-site specific basins. I used the ANA coded data instead of performing our own delineation

using a DEM because first, this is what the ANA has already done, and second the study region of interest is relatively flat, which can make a straightforward DEM-based basin classification difficult without local knowledge, especially with relatively coarse DEM data.

River Basin Delineation

I aggregated the ANA-coded basins to a point of interest along a river (a process hereafter referred to as 'basin delineation') for flow gauge sites that met certain criteria: i) the flow site was located in the custom study region, ii) the flow site was unflagged, iii) the flow site was not at or immediately downstream of a reservoir, and iv) the flow site had at least one year of data between 2000 - 2012 (which may or may not include missing values - the criteria was based on start and end date alone). (The requirement for one year of data between 2000-2012 was specific to research that motivated the creation of the data package. If the date range is excluded from selection criteria, the basins of an additional 100 sites can be delineated.) The number of flow sites that met this criteria out of the original 744 is 377. When the basins are subset further to those with no flag, no onsite reservoir, and with <10% basin area impacted by a reservoir, then the number of basins is 332. These are the basins shown in Figure 2.4, however all 377 are included in basin boundary GIS files. Figure 2.4 shows selected flow site drainage basin locations entirely within the study region.

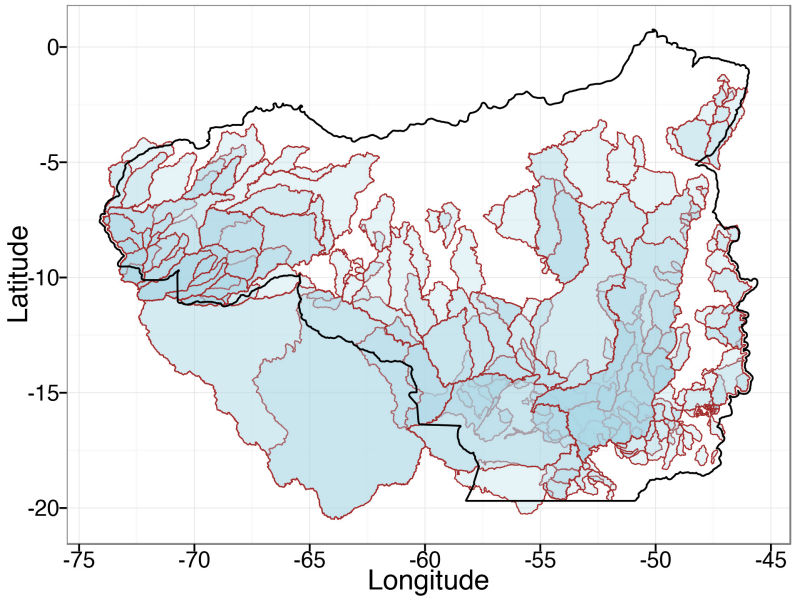

Figure 2.4: Drainage basins corresponding to flow gauge locations

Basins are shown as light blue polygons in and around the custom study region (black line), and transparency is used to indicate overlapping basins. Data source: ANA Geospatial Metadata Portal [37].

The basin delineation algorithm aggregates basin units from the ANA-coded data into complete basin areas that drain into the selected flow gauge sites according to the numbering system assigned by the ANA; I aggregate all basins identified as being upstream of a 'root' basin that overlaps with the flow gauge site. Each delineated basin was visually inspected for accuracy (plot of basin boundary, root basin, flow gauge site location, and rivers within basin boundaries), and modifications to the basin aggregation algorithm were made where needed; many basins required customized sub-basin additions within the algorithm. The script used to carry out basin delineations is not included with this data package (due to the fact that the format of the ANA-coded data has changed since our acquisition of those data - see Section 2.2), however the ANA data that I used (downloaded in 2014), as well as the script, is available upon request. The flow basin delineations allows for calculation of flow site drainage areas (enabling area-normalization of flow time series, for example). Additionally, nested basin groups were identified using basin area overlap, and were assigned a numeric indicator value in the flow QC notes.

Reservoir Basin Delineation

Reservoir delineations were completed using the same method used to delineate flow basins (described above), except that basins were aggregated to the area upstream of the reservoir locations (see Section 2.2) instead of flow site locations. The reservoir basin delineations allowed us to calculate the percentage of a flow gauge site's upstream area that drains into (is impacted by) a reservoir. Ultimately, basins with a threshold percentage (e.g. 10%) of their upstream area impacted by a reservoir may be excluded from analysis.

2.4 Conclusion

Documentation of these data acquisition and processing steps was required in order to justify use of a novel curated dataset for analysis, and is intended to provide detailed information to other future users of data acquired from ANA and ANEEL. Raw through quality controlled data and associated records, including code, GIS and time series CSV data files, and QC notes for custom data filtering are included in the Figshare [31] instance of "Curated rain and flow data for the Brazilian rainforest-savanna transition zone", and an analysis-ready version is discoverable through the CUAHSI Hydroshare platform [32]. Code used to carry out the analyses described (other than code already provided in the data package - see the "ReadMe" file located at the Figshare instance of the data package) is available upon request.

Chapter 3

Addressing rainfall data selection uncertainty using connections between rainfall and streamflow

3.1 Introduction

Quantifying precipitation patterns at regional scales is essential for water security [58, 59], but is compromised by discrepancies in rainfall datasets [60, 61, 62]. Spatial rainfall data products have proliferated, drawing on differing information sources, using different techniques to impute that information through space, and varying in their spatial extent and spatio-temporal resolution [63]. The proliferation of such rainfall datasets facilitates applied research at regional spatial scales, but raises the risk that naïve use of an individual rainfall product may introduce bias into subsequent analyses, relative to the full range of representations of the rainfall field available [64]. Addressing this risk requires quantifying the differences between available rainfall data products, and, if possible, identifying and working with only those datasets that are most suitable for the intended analysis. Here we firstly show that the differences across daily rainfall datasets, for a test case in Northern Brazil, are large enough to require such uncertainty characterization. Next we demonstrate that comparison of datasets with a mechanistically related, but independently observed environmental variable, in this case streamflow, can provide a basis for selecting among available rainfall products. Although our proximate goal is to identify and reduce the uncertainties associated with naïve selection of a rainfall data product for hydrologic purposes, the approach is generalizable to other climatic products and applications.

Regional rainfall data are collected through remote sensing (RS) and in-situ (IS) rain gauge observations. At regional scales, and in remote, rural or developing regions, the rainfall data products generally available and most applicable for hydroclimatological analyses [61] are based on RS data, IS data, or both. IS data provide precision and accuracy at a point, but are often distributed sparsely and heterogeneously in space, and discontinuously

in time [65, 66], and may pose quality control challenges [67, 68]. RS data have consistent coverage and represent spatial heterogeneity, but are often biased, with uncertainties that are dependent on topography, climate, and the level of spatial and temporal aggregation [60, 69, 62]. Differences between rainfall datasets emerge, especially at daily or sub-daily temporal resolutions [64], mostly due to artifacts introduced during data processing. For RS data, such artifacts can include a combination of satellite data retrieval technologies and associated processing algorithms, as well as IS calibration sources and methods [61]. For IS data, artifacts may derive from gauge measurement quality, availability, and the imputation and/or interpolation methods used [70]. While RS data may be a preferred alternative to IS data in settings with sparse rain gauge networks [71], at regional scales, both data types, and their spatial imputations, are expected to differ from 'true' (and unknown) rainfall fields.

Consequently, the challenge of data selection given the uncertainty associated with datasets is not to determine the 'most accurate' dataset, for which there is no universal assessment [61, 72], but instead to quantify the uncertainty in any given analysis that derives from the different representations of reality by the available ensemble of data products. If possible, data selection should also identify the most 'fit-for-purpose' dataset, based on its fidelity to the features of rainfall (e.g. mean, extremes, trends, or correspondence with an independently measured and mechanistically related environmental variable) most pertinent to a given study topic.

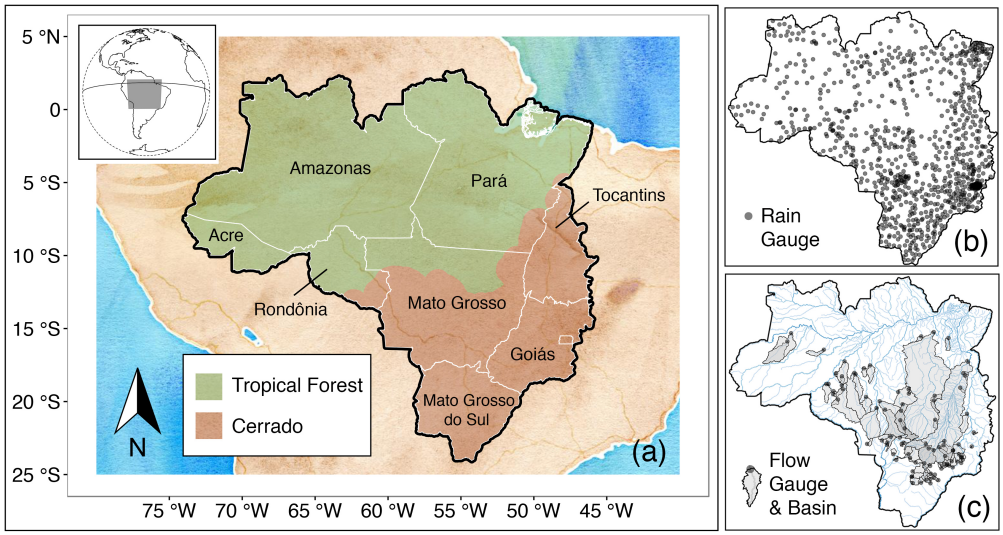

Figure 3.1: Study region and locations of in-situ (IS) rainfall and streamflow gauges

Panel (a) shows the Amazon-Cerrado transition states of Brazil. Panels (b) and (c) show the location of 942 rain gauges and 89 streamflow gauges and associated catchment areas (river basins) used in the analysis. The majority of the river basins in (c) drain to the north, and are headwater basins of the Amazon and Araguaia-Tocantins River Basins; basins located in the south are headwater basins of the Paraguay and Paraná River basins. These maps were generated in R, Version 3 (https://cran.r-project.org/) [73] using data from the curated data package accompanying this manuscript [32] and biome boundary data[74].

Our case study region, the rainforest-savanna (Amazon-Cerrado) transition zone in Brazil (Figure 4.1 (a)) has experienced dramatic changes in land-cover, with anticipated feedbacks to regional climate [75, 76], and thus to the wide variety of rainfall-dependent ecosystem services provided in the region, including agricultural industries, the hydropower sector [18, 77], and extensive regional forest. Variability and change in the Amazon and surrounding region's precipitation therefore affect Brazilian economic, food, and energy security, and potentially also the health of the Amazon rainforest and the global climate system [1, 78, 79]. Rainfall in center-west and northern Brazil is monitored through a relatively sparse rain gauge data network (15 or fewer rain gauges per 10^4 km^2), comparable to inland regions of South America; sub-Saharan Africa; and central, east, and southeast-Asia [80]. These low densities are likely to result in non-trivial differences between regional rainfall data products (in Switzerland, rain gauge densities of > 24 rain gauges per 1,000 km ^{2} were required to avoid density-dependent biases [66]).

Rainfall data in center-west and northern Brazil are therefore likely to be inaccurate at regional scales, yet remain highly relevant to a wide range of policy and planning efforts. For the purposes of this paper, we focus on the quantification of regional daily rainfall statistics needed for hydrologic analyses. Daily rainfall data, or statistical representations thereof, are needed as input to a broad range of hydrological models and empirical analyses that assess spatial or regional trends or drivers of flow variability [81, 82]. We analyze a suite of statistical descriptors of daily rainfall, including the daily mean rainfall depth, wet-day mean rainfall depth, and percent occurrence of wet days. These all influence streamflow response [83], and are referred to as "rainfall characteristics" in the remainder of this paper (results for a more expansive range of rainfall statistics are also presented as Supplementary Information).

The rainfall datasets used in this analysis (Table 3.1) include four global and quasi-global gridded (RS and IS) products, and five custom interpolations of the Amazon-Cerrado rain gauge network, containing 942 gauges (Figure 4.1 (b)) and managed by the the Brazilian government water management agency (Agência Nacional de Águas - ANA). The curated IS rainfall and streamflow data used in this analysis are provided in a data package: "Curated rain and flow data for the Brazilian rainforest-savanna transition zone" [32]. We interpolated each day's set of reporting rain gauges over a 16 year period, from January 1, 1998 to December 31, 2013, using five interpolation methods ranging from a naïve nearest-neighbor to more sophisticated geostatistical approaches (see Methods).

Intercomparison of these products is not straightforward. Point (IS) estimates of rainfall are not directly comparable with gridded (RS) estimates [84, 85]. Because streamflow responses arise at river-basin scales, we focus here on an intercomparison at spatially averaged river-basin domains, calculating the rainfall characteristics over 89 river basins in the study region (Figure 4.1 (c)), as well as on a 0.25° resolution grid. Given this focus, and the characteristics of the region and its rainfall, we might expect gridded RS products to be preferred. RS products are often preferred over IS products in regions where low gauge density prohibits high quality interpolation [66], and the flat, low-altitude, and moderately wet conditions in central and northern Brazil are considered optimal for RS rainfall retrieval

[86, 60, 87, 62].

Our approach to data selection and quantification of uncertainty involves an initial intercomparison of the rainfall characteristics, at grid and basin scales, across the nine rainfall datasets. In the absence of an independent set of empirical measurements against which to compare the datasets, the resulting range in the rainfall characteristics across datasets provides an ensemble measure of the uncertainty associated with these characteristics, which we measure using the maximum absolute deviation (MAD) and standard deviation across datasets for each statistical measure in each basin. To illustrate how such dataset differences may propagate into subsequent analyses, we compute several hydroclimatic indices or analytical results - the runoff ratio (ratio of annual runoff to rainfall), the evaporation ratio (ratio of annual evapotranspiration to rainfall), the Horton index [88] (ratio of evapotranspiration to available soil water), and long-term (inter-annual) trends in daily rainfall, evaluated on monthly timescales for each basin, and again compute MAD and the standard deviation for each basin. The range in these computed indices and trends provides an ensemble description of hydroclimatic uncertainty due to the propagation of data selection uncertainty into these simple analytical outputs.

Having demonstrated that the differences in rainfall characteristics and their propagation into simple analyses are large enough to cause concern, we next attempt to select a rainfall dataset for use in hydrologic studies, by the approach of comparing rainfall datasets to an independently measured, but mechanistically related, environmental variable. In this case, we use streamflow records across 89 river basins to provide such an independent metric. Given the mechanistic connection between streamflow and rainfall, whereby preceding rainfall events drive subsequent streamflow increases, we use measures of time series correspondence or similarity between daily rainfall (at river basin scales) and streamflow for this intercomparison. Specifically, we treat datasets that maximize the correlation between rainfall and streamflow timeseries, and the correspondence of rainfall with streamflow peaks (see Methods), as being the most informative for hydrologic studies.

3.2 Results

Rainfall characteristics

Figure 3.2 (a) shows mean daily rainfall over the study period (1998-2013) at individual grid cells for all nine rainfall datasets, demonstrating relative consistency in large-scale spatial patterns and magnitudes of rainfall, although the mean at individual locations can differ substantially. Figure 3.2 (b), however, demonstrates dramatic differences in the representation of wet-day ($\geq$ 1mm/day) rainfall, illustrating, for example, that rainfall detection (to which RS data errors are principally attributed [60]) and representation of extremes differentiate datasets. Differences across datasets - in spatial patterns and magnitudes - persist across a suite of other statistics (see Supplementary Figures 2-5, which depict grid-cell-level median and wet-day median rainfall depths, maximum and standard deviation of rainfall

Table 3.1: Daily rainfall datasets

Category	Name	Description	Type	Resolution
Gridded	GPCP	Global Precipitation Climatology Project (GPCP), Version 1.2	RS, IS	1°
	CPC	Climate Prediction Center (CPC) Unified Gauge-Based Analysis of Global Daily Precipitation, Version 1 and RT	IS	0.5°
	TRMM	Tropical Rainfall Measuring Mission (TRMM) 3B42, Version 7	RS, IS	0.25°
	PERSIANN	Precipitation Estimation from Remotely Sensed Information using Artificial Neural Networks - Climate Data Record (PERSIANN-CDR), Version 1.1	RS, IS	0.25°
Custom	UKP	Universal Kriging with PERSIANN (predictors: elevation, latitude, longitude, and PERSIANN-CDR)	IS, RS	0.25°
	UK	Universal Kriging (predictors: elevation, latitude, longitude)	IS	0.25°
	OK	Ordinary Kriging	IS	0.25°
	IDW	Inverse Distance Weighting	IS	0.25°
	VP	Voronoi (or Thiessen) Polygons	IS	0.25°

Gridded data are global and quasi-global products; custom data are regional (local) interpolations of IS rainfall data obtained from Brazil's Agência Nacional de Águas (ANA). See Methods for sources and additional details.

depths, mean annual total rainfall, and wet-day occurrence of rainfall). Figure 3.2 shows point-estimates of rainfall properties at individual grid cells. However, we are primarily concerned with observations of area-integrated rainfall, and the remaining results pertain to areal spatial units (either sample areas or river basin units, as noted).

Figure 3.3 shows variation in the same statistics as presented in Figure 3.2 - the mean and wet-day mean, as well as the occurrence of wet days, over river basin units of analysis (Supplementary Figure 6 shows basin-level percentiles, mean annual total, standard deviation, and maximum). Again, there is overlap in the mean daily rainfall estimates, but significant variation in wet-day mean values across rainfall datasets. These results suggest that the rainfall datasets can be divided into two groups: the first (I) includes the gridded datasets GPCP (RS), CPC (IS), and TRMM (RS), and the nearest-neighbor interpolation VP (IS); and the second group (II) includes the remaining interpolations UKP (IS, RS), UK (IS), OK (IS), and IDW (IS), and the gridded PERSIANN (RS) dataset. Figure 3.3 shows that group II datasets report a greater number of wet days, but lower mean rainfall on those wet days, relative to group I. The lower mean wet-day rainfall of group II stems from the fact that group I data report more wet-day extremes (see Supplementary Figures 3 and 6), which upwardly bias the mean wet-day rainfall of group I, despite those data showing fewer wet days. While group I *wet-day* medians are also greater than group II in accordance with wet-day means, group I *all-day* medians are *less* than those of group II (see Supplementary Figures 2

and 6). This is due to the combination of greater wet-day occurrence and medium-intensity rainfall (1-10 mm/day) in group II (see Supplementary Figure 7). These differences persist across the range of rain gauge densities in the study region (see Supplementary Figure 7).

Greater wet-day occurrence in group II custom interpolations (UKP, UK, OK, and IDW) likely results from greater rates of local detection of medium intensity rainfall by rain gauges relative to satellite sources, combined with spatial smoothing of those rainfall events. In the case of PERSIANN, elevated wet-day occurrence can be attributed to a combination of the rainfall estimation algorithm and/or incorporation of multiple RS and IS rainfall products that are unique to this RS dataset compared to earlier RS products (GPCP, CPC, TRMM) [89]. In summary, divergent features of the group I datasets (lower wet-day occurrence and medium intensity rainfall depths, greater extremes), and group II datasets (greater wet-day occurrence and medium intensity rainfall depths, lower extremes) may result in similar mean daily average values across large regions as shown in Figure 3.2 (a). Thus, there may be consistency across rainfall datasets in analyses relying upon regional mean daily rainfall values (only). However, different datasets will propagate significant uncertainty into analyses relying on estimation of wet-day rainfall occurrence or depths, quantiles, and extremes.

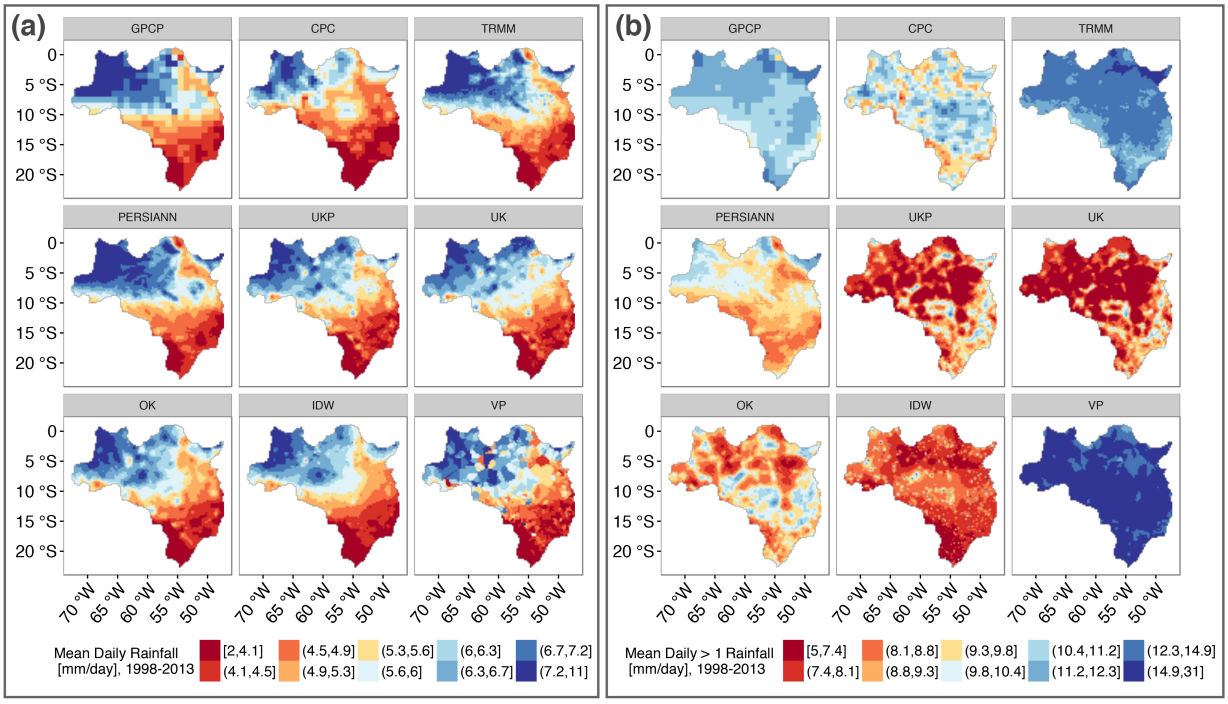

Figure 3.2: Spatial variation in the representation of descriptive statistics by different rainfall datasets

Panel (a) shows mean daily rainfall (depths in mm/day), and panel (b) shows mean wet-day rainfall (depths in mm/day for days with ≥ 1 mm/day). Both statistics were calculated at each 0.25° resolution grid cell in the study region using all daily data between 1998-2013. See Table 3.1 and Methods for dataset details. GPCP, TRMM, and PERSIANN are gridded datasets comprised of RS and IS data sources; CPC is a gridded dataset comprised of IS data; UKP is a custom interpolation of IS and RS (PERSIANN) data sources; UK, OK, IDW, and VP are custom interpolations of IS data. These maps were generated in R, Version 3 (https://cran.r-project.org/) [73].

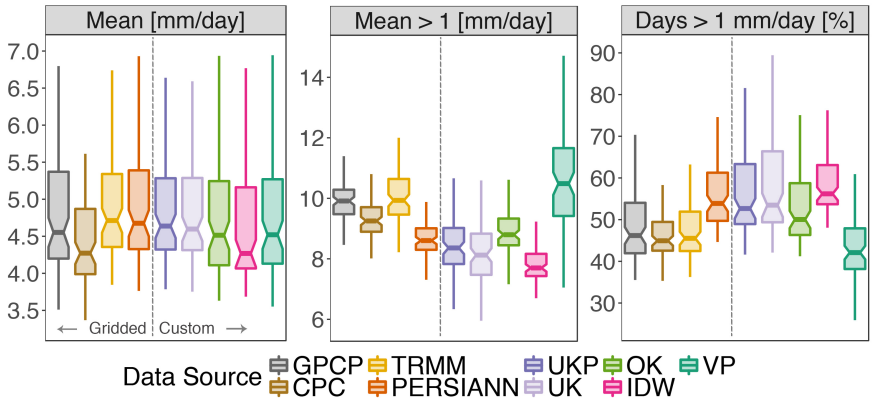

Figure 3.3: Daily rainfall statistics in river basins according to different rainfall datasets

From left to right, respectively, the panels show the simple daily mean, wet-day ($\geq$ 1 mm/day) mean, and wet day occurrence (percent of days with $\geq$ 1 mm/day) of rainfall for 89 river basins. Each boxplot is generated with n=89 (river basin) statistic results, calculated using basin area-average rainfall from the given rainfall dataset (colors) from all days between 1998-2013. Outliers are not shown. The vertical dashed line separates gridded from custom datasets (see Table 3.1).

Calculation of the maximum absolute deviation (MAD) between any two datasets' area-average rainfall (averages over areas of the 0.25° grid) provides a simple quantification of dataset divergence and thus the range of the data ensemble. We calculated 1998-2013 MAD at daily, monthly, and annual time scales at 100 regularly-sampled locations, for areas ranging from large to small (circles with radii of 200 km and 10 km, centered at the same 100 locations). This sample design accounts for the fact that different regions, and differently-sized sample units, have different rain gauge densities. At a daily resolution, the mean (median) MAD between any two datasets' area-average rainfall is 7-12 mm (5-8 mm); at a monthly resolution, it is 56-97 mm (46 - 82 mm); at an annual resolution, it is 372-576 mm (310-497 mm). The ranges are from statistics calculated for the large to small sample units, respectively. These differences are comparable to global climate model (GCM) biases: projections from the Coupled Model Intercomparison Projects Phase 5 (CMIP5) have annual biases relative to a single rainfall data product of -25% (approximately -250 to -550 mm/year) in northern Brazil [90], indicating that selection of a different rainfall dataset for reference has the capacity (at an extreme) to either eliminate or double estimated model bias.

Trends and Hydroclimate Indices

Evaluation of hydroclimatic indices and temporal rainfall trends demonstrates the propagation of rainfall data selection uncertainty into a standard analysis. Although temporal trend analysis is not especially meaningful over a 16-year time period, it demonstrates the potential for trend detection and attribution to be amplified or eliminated by data uncertainty. We calculated monotonic trend slopes (corrected for monthly correlation) and associated p-values for total rainfall by month for all 89 river basin in the study region between 1998-2013 (see Methods). Variation in the estimated trend slopes for basins where at least one rainfall dataset had a statistically significant trend are shown in Figure 3.4. Trend slopes, particularly for basins in the north of the study region where rain gauges are especially sparse, do not agree across rainfall datasets. Rainfall datasets agree on the sign of the slope in only eight of the 24 basins (four basins with all positive slopes, and four basins with all negative slopes).

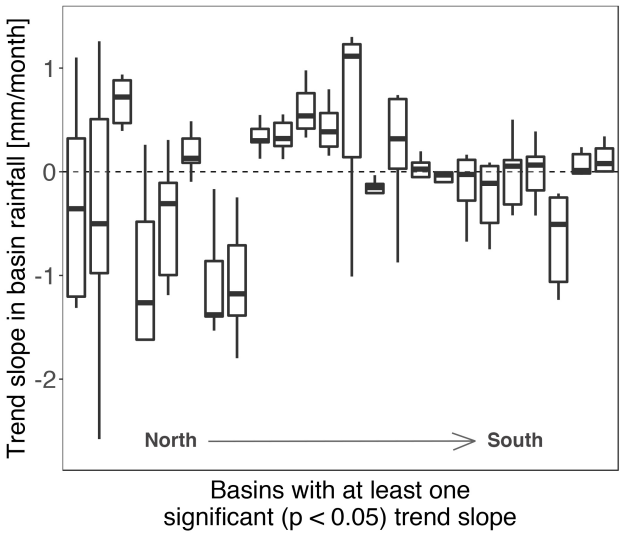

Figure 3.4: River basin monthly total rainfall trend slope variation across rainfall datasets

Each boxplot shows the distribution of individual basin trend slope coefficients (excluding outliers) estimated using nine rainfall datasets; the 24 basins shown are those for which at least one rainfall dataset has a significant (p ¡ 0.05) trend; not all trend slopes are significant. Basin boxplots are ordered left to right according to the latitude of the basin centroid.

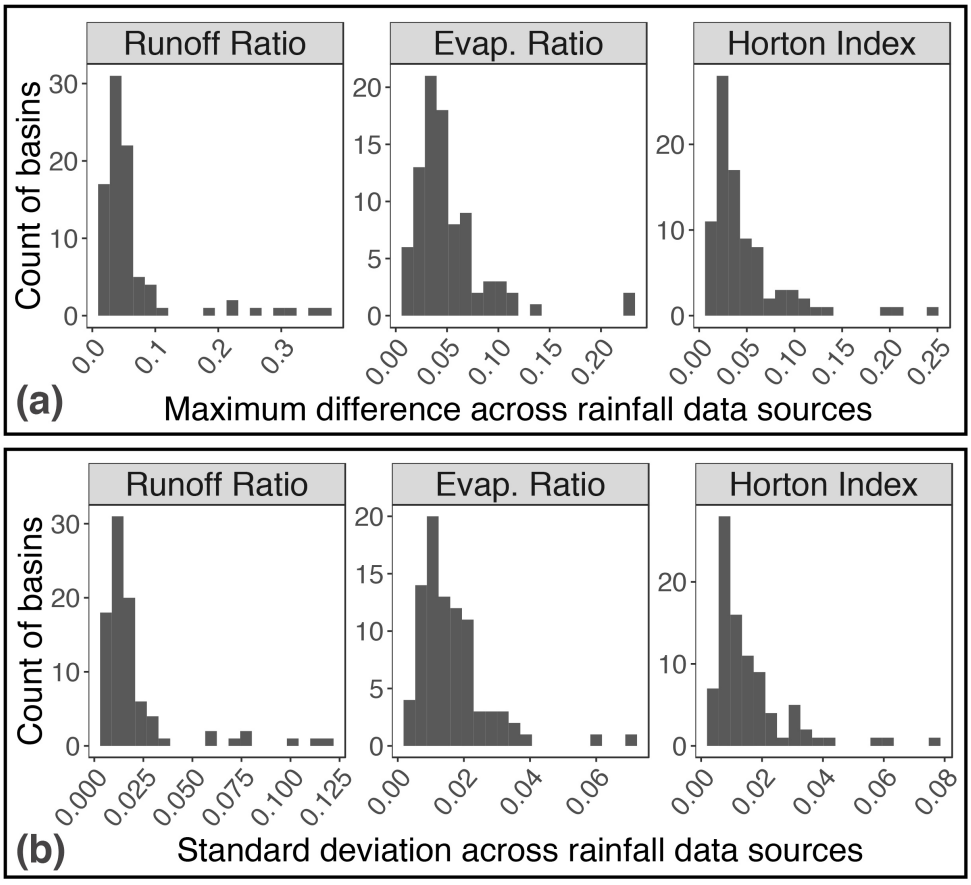

Figure 3.5: Distribution of river basin hydroclimate index differences and variance across rainfall datasets

Panel (a) displays histograms of maximum absolute deviations between individual river basin hydroclimate indices calculated using each of the nine rainfall datasets. Panel (b) displays histograms of the standard deviation of individual river basin hydroclimate indices across the nine rainfall datasets. The runoff ratio is the runoff fraction of rainfall; the evaporation ratio is the evapotranspiration fraction of rainfall; and the Horton index is the evapotranspiration fraction of available soil water (see Methods for details). All index values range between zero and one.

The propagation of rainfall data selection uncertainty is further illustrated by hydroclimatic index measurements made using the different rainfall datasets. Hydroclimatic indices provide information on the relationships between climate, land use, and hydrology, which are critical to the examination of land use and climate change [91, 88]. They are estimated using both rainfall and streamflow at river basin scales (see Methods). The runoff ratio is the fraction of rainfall discharged from a river basin as streamflow (as opposed to evaporated or transpired at the land surface, or percolated to deep groundwater); the evaporation ratio complements the runoff ratio - it is the fraction of rainfall evapotranspired (as opposed to discharged or percolated); the Horton index compares evapotranspiration to soil water stores (as opposed to total rainfall). The mean (median) maximum absolute deviation between basin-level index values generated using any of the nine rainfall datasets (see Figure 3.5 (a)) is 0.05 (0.04) for the evaporation ratio and Horton index, and 0.06 (0.04) for the runoff ratio; the difference exceeds 0.25 - a quarter of the entire index range - for some basins. Similarly, the mean (median) standard deviation of basin-level index values (see Figure 3.5 (b)) is 0.02 (0.01) for all three indices; and can exceed 0.05 in some basins. Streamflow data is the same for all calculations within each basin, so these results demonstrate the sensitivity of basin-scale analyses to rainfall input data alone.

In the absence of information on a 'best' rainfall data source, and knowing that data selection uncertainty will propagate into analyses as demonstrated in Figures 3.4 and 3.5, distributions of index values obtained from multiple rainfall datasets can be used to quantify data selection uncertainty. For example, the mean of the standard deviations across all basins for a given index (e.g. mean of values shown in each panel of Figure 3.5 (b)) may be treated as an index- and region-specific standard deviation (s) attributable to rainfall data selection uncertainty. According to our analysis, in rainforest-savanna transitional Brazil, s is approximately 0.02 for all three indices. A straightforward confidence interval for the mean of index values obtained using the nine rainfall datasets over an *individual* basin ($\bar{x}$) in our study region is: $CI = \bar{x} \pm z^* SE_{\bar{x}}$, where $SE_{\bar{x}} = s/\sqrt{89} = 0.002$ (see Supplementary Discussion for further details).

Rainfall and Streamflow Correspondence

Figures 3.2-3.5 demonstrate the need for a procedure to guide rainfall data choice prior to conducting analyses. We build on the precedent for evaluating rainfall data quality using the correspondence between rainfall and river flow [86, 92, 71] by measuring the empirical correspondence between rainfall and streamflow records using two performance statistics: non-parametric Spearman's rank correlation, and peak correspondence - the rate at which distinct rainfall peaks correspond to distinct flow peaks within a basin-specific response time window (see Methods). Streamflow rises and peaks in unregulated, rain-fed rivers are caused by preceding rainfall events in the rivers' catchment, so the correspondence between appropriately-lagged and basin-integrated rainfall, and basin streamflow, measures a rainfall dataset's ability to capture area-integrated rainfall patterns.

In validation tests, rainfall data from seven Australian river basins was randomly perturbed using additive noise, and true and perturbed rainfall datasets were evaluated relative to streamflow using the performance statistics. Both performance statistics identify the correct rainfall dataset 100% of the time when the random noise is equivalent to or greater than basin rainfall standard deviation. In cases where random noise is less than or equal to half the basin rainfall standard deviation (when differences between datasets are small), correlation still identifies the correct rainfall dataset 100% of the time, however peak correspondence identifies the correct dataset on average (across the seven test basins) 79% of the time or less (see Supplementary Discussion for details). Specifically, peak correspondence performs perfectly (100% correct identification) in some basins, but not others, when the signal to noise ratio is low. This is likely due to peak correspondence's reliance on quick (storm) runoff response signatures (see Methods), which may vary in quality across different basins. In the study region, the 1998-2013 average maximum absolute deviation (MAD) between any two rainfall datasets on a daily time scale is 7-12 mm; the range of grid cell-level (temporal) standard deviations averaged across the study region for each individual rainfall datasets is between 7-13 mm. Thus, in the study region, individual rainfall dataset variation is on the order of variation between datasets, indicating that peak correspondence will perform as well or nearly as well as correlation in identifying datasets with greatest correspondence to flow. This is confirmed by the similarity in results from both statistics for the Brazilian data.

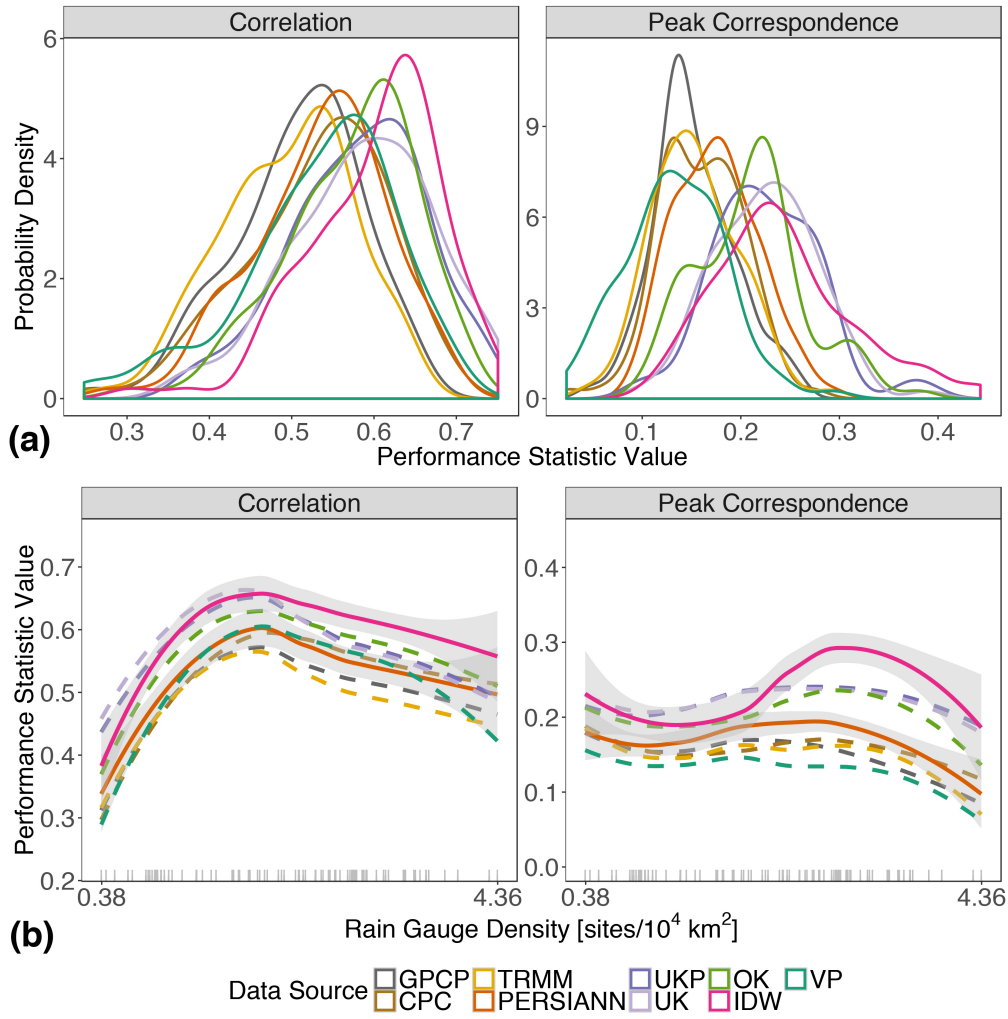

Figure 3.6: Differences in rainfall data quality as indicated by performance statistics measuring correspondence with streamflow

Panel (a) shows kernel-smoothed empirical probability distributions of performance statistics (correlation and peak correspondence) by rainfall dataset. Panel (b) shows the same performance statistics plotted as local regression-smoothed curves across the range of rain gauge densities in the study region, with 95% uncertainty intervals (shaded). In (b), solid lines and shaded regions indicate the best-performing gridded (PERSIANN) and custom-interpolated (IDW) datasets and their 95% uncertainty intervals, respectively; non-overlapping uncertainty intervals indicate distinguishable performance between datasets; dashed-lines indicate all other datasets with uncertainty intervals that are not displayed, but are of similar width; gray tick-marks at the bottom illustrate the spread of rain gauge densities in the 89 river basins.

Figure 3.6 presents distributions of the performance statistics in 89 river basins (panel a), and illustrates the sensitivity of the performance statistics to rain gauge density within the river basin (panel b). The better the performance of the dataset, the farther to the right are

the masses of the distributions in (a), and the higher the curves are in (b). We found that custom interpolations of IS data using IDW and kriging (UKP, UK, and OK) out-performed the gridded datasets for both performance statistics, with IDW performing best overall. In agreement with these results, equivalent or superior performance of the IDW method relative to other interpolations including kriging and VP, specifically for hydroclimatological applications, has been observed in other regions as well [66, 93]. The best performing gridded dataset is PERSIANN, whose statistics in the study region more closely resemble those of custom interpolated datasets than other gridded products. The differences between performance statistic distributions are statistically significant (as evaluated by non-parametric two-sample Kolmogorov-Smirnov tests, see Supplementary Table 1), consistent across gauge densities (as illustrated in Figure 3.6 (b)), as well as consistent across location (as indicated by latitude) and basin size (see Supplementary Figure 8), and season (see Supplementary Figure 9). The rain gauge densities in (b) are 1998-2013 averages of basin-area daily densities according to the IS data; they do not directly pertain to gridded datasets, but they are indicative of gridded dataset input gauge densities because gridded product source data (used directly, or for calibration) also comes from Brazilian government agency sources.

3.3 Discussion

The 'data selection uncertainty' problem identified here is similar to the 'gigo' (garbage in, garbage out) problem in modeling, but applied to regional data analysis. Although the need to base analyses and interpretation on high quality data appears self-evident, the inability to directly observe the true spatial process of interested at regional scales, and thus to *a priori* discriminate between a wide array of available or self-generated regional data products, means that regional data selection is not trivial. Instead, it should motivate environmental scientists to consider the state of practice in the field, with respect to the use of, and confidence placed in, the use of regional climatic data products. For example, in Northern Brazil, where we have identified significant and meaningful differences between rainfall datasets, a wide range of studies draw inference about historical climate patterns and trends [57], drought [78], the effects of land use change on hydrology [2, 9], and relationships between hydroclimate and agriculture [94, 95], *without* confronting data selection uncertainty. Our analyses suggest that the conclusions of these studies must be treated with caution, as the magnitudes of difference or trends within data products may be comparable to the magnitudes of difference between data products. Several studies in the region do explicitly addressed data selection uncertainty: by correlating rainfall and streamflow datasets and selecting the rainfall product with the greatest correspondence [4], and by demonstrating that multiple rainfall products would generate similar results [96]. Overall, however, data selection uncertainty remains inconsistently acknowledged and unaccounted for by practitioners.

The empirical time series and signal-processing methodology used here (i.e. performance statistics) offers an approach to evaluate rainfall data quality for hydrological purposes across

multiple river basins and at large spatial scales and is arguably an improvement on the state of practice for regional hydrology. Traditional rainfall data error estimation frameworks infer rainfall data quality at points using cross-validation methods, or over river basin areas based on runoff predictions made via a model [86, 92]. Point-scale evaluations do not address areal-scale data quality, and at regional scales - i.e. the 89 basin region in this study - a model-based approach would require 89 separate runoff model calibration/validation procedures, and would not generate results that are comparable between basins because the calibration error would be unique for each basin [67, 71]. Furthermore, the attribution of prediction error to calibration would be confounded with rainfall data input uncertainty. Lastly, the quality and reliability of rainfall-runoff model prediction relies on input stationary [97, 98], which is not guaranteed in the study region due to climate and land use change. Thus, model-free approaches are desirable. Our empirical approach capitalizes on the relationships between variables (rainfall, streamflow) rather than on their exact values to evaluate rainfall dataset quality at basin scales. This method complements standard model-based evaluation, but is scalable and generalizable over large regions that challenge the use of models.

While it was possible to identify a best performing rainfall dataset based on streamflow correlation in this region, the results are likely to be site specific and specific to applications in which comparing rainfall signals to streamflow signals offers an appropriate test of quality. Evaluations should be made separately for new study areas, and potentially by comparison to reference datasets other than streamflow for different study purposes. For example, streamflow intercomparisons would not necessarily inform the suitability of a rainfall dataset for surface soil moisture estimation purposes, as would microwave remote sensing data. Similarly, interpolation methods such as UK (which can control for elevation) would likely improve upon IDW in mountainous areas. The differences between the datasets' performance statistics were reduced when data were aggregated or smoothed over time, consistent with previous studies that have shown RS data to correspond well to IS data with greater temporal aggregation [99, 100]. Thus, at coarser temporal resolutions (monthly, annual), convenient gridded products remain attractive.

Critical climate change adaptation decisions are likely to derive from the understanding of emerging trends and variability in regional rainfall estimates. These results highlight the often-unacknowledged problem of 'data-selection uncertainty' in the detection and attribution of environmental change [101, 102], and demonstrate a need for increased effort in quantifying this uncertainty and justifying data choice because analysts may reach divergent understandings due to data selection alone [103]. Identifying the often weak signals of change in noisy datasets is challenging, but analysts can reduce the uncertainty derived from data choice by (i) justifying dataset choices using selection methods such as the performance statistics demonstrated here, and/or (ii) including estimates of data-selection uncertainty (e.g. confidence intervals) in their findings. Evaluation of rainfall data prior to hydroclimatological analysis is both feasible (if streamflow records are available) and necessary. In contrast to the use of climate model outputs in analyses - where characterization of an ensemble of equally uncertain projections is best practice - if an individual dataset corresponds more closely with a reference of choice (e.g. streamflow) than other datasets, that dataset

should be used for analysis.

3.4 Methods

Data

Gridded datasets include: Global Precipitation Climatology Project (GPCP) Version 1.2 [104]; Climate Prediction Center (CPC) Unified Gauge-Based Analysis of Global Daily Precipitation Version 1 and RT data [105, 106]; Tropical Rainfall Measuring Mission (TRMM) 3B42 Version 7 [107]; and Precipitation Estimation from Remotely Sensed Information using Artificial Neural Networks - Climate Data Record (PERSIANN-CDR) Version 1.1 [108, 89]. GPCP, TRMM, and PERSIANN were acquired from public repositories via the IRI/LDEO Climate Data Library [109] and CPC from the `raincpc` R package [110]. IS rainfall, streamflow, and geographic information systems (GIS) data were acquired from the Agência Nacional de Águas (ANA), and reservoir locations (used to select only unregulated river basins for analysis) from the Agência Nacional de Energia Elétrica (ANEEL). Of a total of 1,171 usable rain gauges in the study region, 942 were active (for varying durations) during the study period, and were used for analysis. Daily streamflow data was obtained for basins fully contained in the study region, gauged for at least a year, and with $< 10\%$ of their area impacted by reservoirs. Analysis was based on 89 basins that met data quality criteria and overlapped with the interpolated rainfall region. All IS rainfall, streamflow, GIS data, and comprehensive documentation on data acquisition and quality assurance/quality control are provided in the "Curated rain and flow data for the Brazilian rainforest-savanna transition zone" data package [32]. Interpolated rainfall data are available upon request from the corresponding author. (See the Supplementary Discussion for more discussion of rainfall data.)

We did not manipulate the spatial resolution of the daily rainfall datasets (all are obtained or generated at 0.25°, except for two sources at 0.5° and 1° resolution - see Table 3.1), nor did we compare the representation of rainfall based on spatial resolution. We did however briefly explore areal rainfall differences across rain gauge densities with respect to the known gauge densities of interpolated IS data - see Supplementary Figure 7). Typically, to compare rainfall datasets, one would aggregate (or disaggregate) rainfall datasets to a common grid using a method that conserves the total amount of rainfall in an area. Effectively, this study aggregates total daily rainfall to river basin units, without modifying original input data; this is done using a grid cell area-weighted mean of all cells located within a basin area, providing an unmodified representation of each datasets' area-integrated rainfall over multiple basin scales, that is both conservative and representative of the practical needs of hydrologists and hydroclimatologists.

Interpolation

We used four common and well-documented interpolation techniques [65, 92, 70, 93]: Voronoi (or Thiessen) Polygons (VP) [111]; Inverse-Distance Weighting (IDW) [112], and Ordinary and Universal Kriging (OK, UK) [113, 114]. All interpolations were done on a 0.25° resolution grid. IDW and OK are local interpolations, for which we set the maximum interpolation distance (radius) to 300 km, an upper bound on estimated mean rainfall correlation distances in this region, which ranged between 100-300 km for IS and RS data, respectively. Rainfall correlation distances were estimated by fitting a semivariogram model [114] to data on a random sample of 1,500 individual days (approximately 1/4 of the days in the full date range), and extracting semivariogram range estimates for each day. UK and UKP are 'universal' interpolations for the study region; their predictions rely on relationships established between predictor variables across the entire study region. Kriging methods can produce negative values, which were set to zero. To avoid edge effects in interpolations, the grid at which rainfall was interpolated is inset from the study region boundary by 100 km (the minimum mean correlation distance). For additional details on interpolation methods, see Supplementary Discussion.

Trends and Hydroclimate Indices

For trend analyses, we used the non-parametric Seasonal Kendall test for monotonic trends in monthly total rainfall with correction for correlation between monthly blocks, and estimated the slope of the trend using the SK slope estimator [115, 116]; these are seasonally-adjusted modifications of the widely-used Mann-Kendall test [117] and Theil-Sen's slope estimator [118, 119] that are targeted to hydrological time series.

Index values were calculated for each water year (October-September) in each basin, using river basin area average daily rainfall depths (mm/day) from all nine rainfall datasets, and streamflow depths (mm/day, which are basin-area normalized volumetric flow rates) at river basin scales. The index values recorded for an individual basin and rainfall dataset combination is the average of annual index values for that basin-dataset combination (there are 15 complete water years between 1998-2013). The runoff ratio (RR) is the simple ratio of total annual (water year) streamflow (Q) to total annual rainfall (P): $RR = Q/P$. Similarly, the evaporation ratio (ER) is the simple ratio of total annual (water year) evapotranspiration ($ET = P - Q$) to total annual rainfall (P): $ER = ET/P$. (Note that in these computations we assumed no deep percolation.) Lastly, the Horton index (HI) is the ratio of evapotranspiration (ET) to available soil water (W): $HI = ET/W$, where soil water $W = P - Q_q$, and Q_q is the direct runoff component of total flow (Q); W is equivalent to the sum of baseflow and ET - the total amount of water accessible to vegetation. Total flow was separated into baseflow and quickflow using a Lynne-Hollick recursive digital baseflow filter (three-pass, default parameter of 0.975) [120]. The Horton index is intended to be calculated over a growing season [88], however, growing seasons vary across the river basins in this analysis, and many are year-round, thus the use of annual data.

Performance statistics

Volumetric streamflow records were area-normalized and separated into baseflow and quickflow (direct runoff) using a Lynne-Hollick recursive digital baseflow filter (three-pass, default parameter of 0.975) [120]; the quickflow component can be more directly compared to rainfall. Both rainfall and quickflow time series were normalized to between 0 and 1. We identified the lag timescale (τ) that maximized the cross-correlation of rainfall and quickflow (the basin response timescale in units of days) for each basin, and lagged rainfall by τ for analysis of correlation and peak correspondence.

With respect to peak correspondence: we classified peaks in the normalized and lag-aligned rainfall and quickflow data by determining the position of peak extrema (observations that are preceded and followed by lower observations), as well as probabilities associated with peaks [121]. The probability associated with a peak quantifies the distinctness of the peak: more significant peaks are those surrounded by *several* lower observations. Peaks with lower probabilities are those that contain more information according to Kendall's information theory [121, 122]. We call peaks with probabilities < 0.05 'distinct' (due to autocorrelation in the rainfall and flow time series, this is not a measure of statistical significance, but may nevertheless be used to distinguish more and less distinct peaks). 'Peak correspondence' is the rate at which distinct peaks in lagged rainfall match those in streamflow over a basin-specific response time window equivalent to $1/4 \times \tau$ (minimum = 1 day) (see Supplementary Figure 10). Correlation between the lagged rainfall and quickflow was assessed using non-parametric Spearman's rank correlation [123, 124]. For more details, see the Supplementary Discussion.

Code availability and computational tools

Code is available upon request from the corresponding author. We carried out all analyses and generated all figures within the Comprehensive R Archive Network (CRAN) [73] programming environment (Version 3) on both Apple and Windows operating systems. See the Supplementary Discussion for a list of utilized software packages.

Chapter 4

Land use change increases streamflow across the arc of deforestation in Brazil

4.1 Introduction

Global river discharge increased at a rate of 0.08 mm/year from 1900 to 2000 due to anthropogenic land use change, accounting for 50-55% of the total increase from all environmental change, while rates increased even faster in South America (0.23 mm/year) [125]. The majority of anthropogenic land use change in South America has occurred in Brazil's agricultural frontier or arc of deforestation, located along the transition from Amazon to Cerrado (tropical Savanna) biomes [1] (Figure 4.1). Forest loss in this region accounted for 41% of global forest loss (53 out of 129 million hectares - Mha) from 1990 to 2015 [126], 70% of which was in the legal Amazon (36 Mha) [127]. Replacement of natural vegetation, including forest and Cerrado woodlands, with pasture and cropland reduces evapotranspiration (ET) [2, 3], which has the primary, direct effect of increasing streamflow [128]. A mix of empirical and model-based studies carried out in select Brazilian river basins demonstrate location- and deforestation scenario-specific increases in stream discharge at daily, monthly, and annual time scales for small to large (10 - 10^5 km^2) basins [129, 4, 130, 5, 7, 16, 8, 9, 27]. Yet, the contribution of historical deforestation to long-term trends in river discharge across the full Amazon-Cerrado regions remain empirically unquantified. Here we provide, for the first time, a large-scale, empirical quantification of the effects of deforestation on long-term trends in river discharge for the Amazon-Cerrado region.

Quantifying land use change effects on flow in the Amazon-Cerrado region is necessary in order to separate the potentially long-lived consequences of forest loss from short-term climatic fluctuations on river flow and the ecosystem services it provides. In center-west Brazil, services include hydropower production, including planned expansions of hydropower generation facilities [24, 25]; navigability of 13,000 km of inland waterways, which transport 45

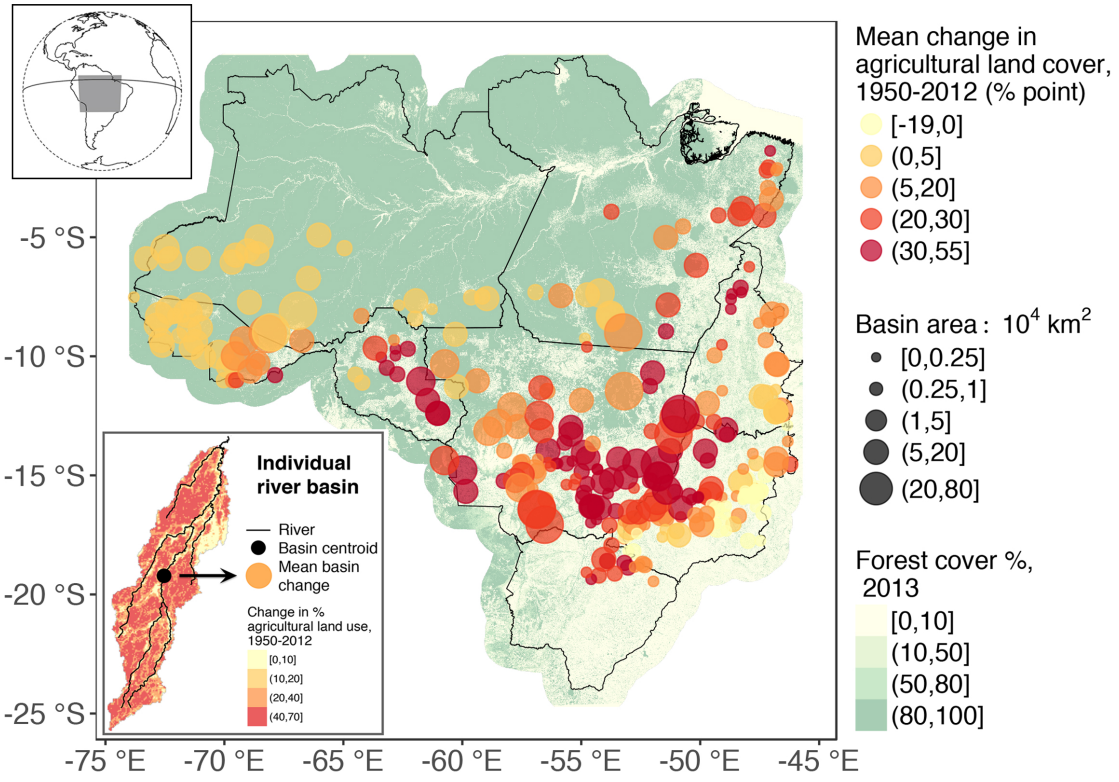

Figure 4.1: Land use change in river basins across Amazon-Cerrado Brazil

Circles indicate the location of 326 river basin centroids [32]; circle colors indicate percentage point change in agricultural land cover [131] between 1950-2012 (calculated from the basin area mean of 1km pixels - see inset); circle sizes indicate the size of the river basin; background shading shows the percentage of land (at 30m pixel-resolution) covered by forest canopy greater than 5m in height in 2013 [132].

million tons/year of agricultural and industrial goods and offer a pathway towards low-carbon regional transport [28]; highly biodiverse aquatic and terrestrial ecosystems, with complex relationships between species' life cycles, flow regime, and river network connectivity [22, 30, 29]; and the potential for agricultural intensification to develop irrigation systems reliant on surface water supplies [1]. The extent to which the hydrologic landscape, including river flow and accompanying nutrient and sediment transport [14, 8], is altered by historical and ongoing deforestation has implications for all such flow dependent processes, as well as understanding of regional and global climate variability and change. Because transport of water vapor to the atmosphere through forest evapotranspiration (ET) drives global atmospheric circulation [133], large-scale deforestation affects local to continental water cycles [134, 135, 136] for which identification of change relies on understanding of streamflow response.

The majority of deforestation in this region between 2000-2013 occurred in medium to high density forests, and the most extreme losses were from dense forests (see SI Text). This implies that river basins in the region (Table B.2) should see significant flow alterations due to

the nature of the vegetation loss, which is primarily from intact forests. There is a discernible relationship between forest cover and streamflow (Figure B.4); however understanding of direct effects of deforestation on streamflow across multiple basins and at regional scales has been limited by system complexity and non-stationarity. Effects are known to vary across basins and are scale-dependent [4, 14, 16]; thus existing studies only hypothesize or suggest regional-scale effects based on individual basin findings. Manipulative experimental studies are limited in scope, scale, and extensibility, and tend to represent extreme or complete short-term vegetation change rather than realistic spatially-heterogeneous, incremental, long-term change relevant to policy and planning efforts [137, 138]. While simulation modeling addresses reliance on the quality of the experimental setting [139], model analyses - by way of data, parameterization and calibration constraints - remain focused on individual basins and deforestation scenarios, and are limited by assumptions and parameter uncertainty [140, 129].

The use of statistical causal inference methods for observational data avoid the shortcomings of previous studies, take advantage of increasingly abundant and accessible environmental data (see Methods, SI Text, and Table B.1), and make it possible to answer a difficult and important scientific question: what is the direct effect of deforestation on streamflow? Unprecedented historical land use change across the Brazilian Amazon-Cerrado transition region presents a unique natural experiment - a setting wherein widespread river basin deforestation is observed rather than assigned by a researcher. An observational study, common in quantitative public health and social science research, is an empirical assessment of the effect of a treatment (i.e. deforestation) when a randomized experiment is not possible, and where complicating factors (i.e. basin physical and climate features) can bias the estimation and attribution of effects [141] (see Figure 4.2). Statistical methods for causal analysis of observational data have a rich theoretical foundation and application history [142], and are beginning to be explored in the environmental sciences [143, 144, 145, 146] due to their suitability in addressing the problem of data-based inference amidst complexity and data (sample) limitations.

Here, we first identify and estimate the relationship between long-term deforestation and average seasonal (monthly) rates of flow using a standard causal difference-in-differences (DID) regression analysis technique, then use a mixed effects statistical model to estimate average annual changes in total streamflow due to deforestation and associated agricultural development. Our approach responds to calls for a "large catchment sample" approach for understanding of hydroclimatological processes at multiple spatiotemporal scales and amidst environmental change [147], and growing awareness within the geoscience community that confirmation of mechanistic model-based understandings in this context, especially for prediction, require causal empirical assessments [148].

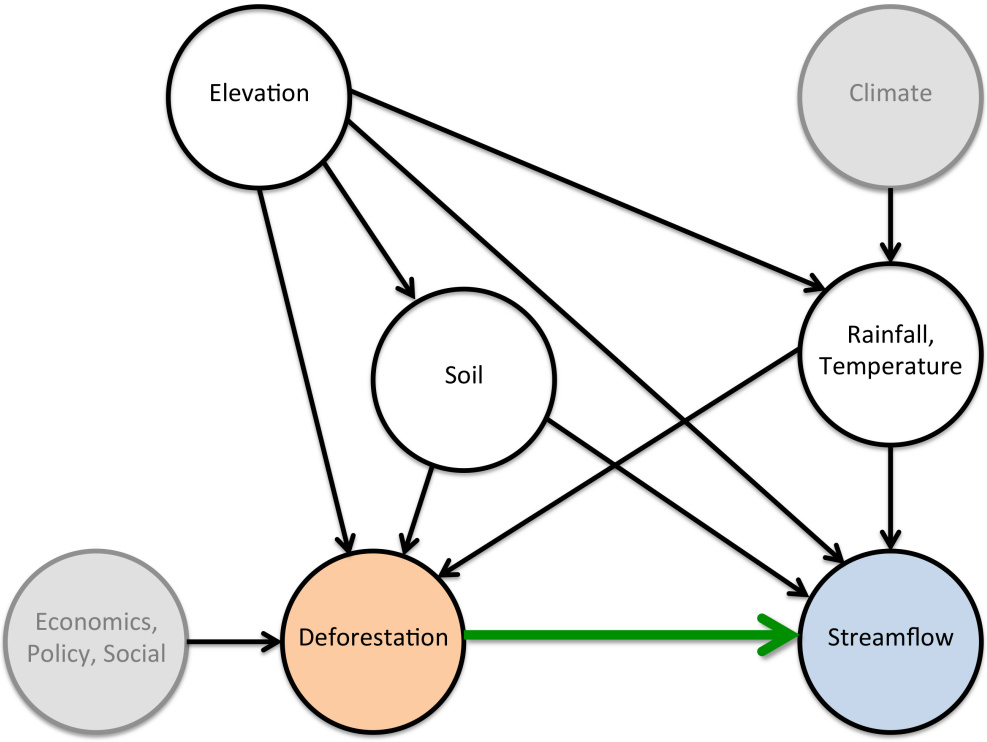

Figure 4.2: Causal diagram, or directed acyclic graph (DAG) for the process governing deforestation effects on streamflow

The diagram shows key variables and relationships between those variables (arrows), and represents a forward-moving process at a single time step. Deforestation is considered a "treatment" variable; streamflow is the outcome variable of interest; elevation, soil, rainfall, and temperature are variables that must be adjusted for (conditioned on) in order to estimate the direct effect of deforestation on streamflow; the variables in gray circles are unobserved (latent) variables. The green arrow indicates a causal path: the direct effect of deforestation on streamflow may be estimated provided the statistical relationship between deforestation and streamflow is adjusted for soil, elevation, rainfall, and temperature.

4.2 Results and Discussion

DIfference-in-differences (DID) analysis

The difference-in-differences (DID) regression modeling approach is an empirical statistical technique that estimates the differential effect of a treatment (i.e. deforestation) on a treatment group versus a control group using observational data in a natural experimental setting [149]. This approach compares flow change outcomes across two periods (before 1990 and after 2000) and two groups of basins (low and high deforestation), and accounts for confounding factors (basin physical and climate features).

The DID model takes the form of a fixed effects log-linear regression (see Methods and SI Text) of normalized flow percentiles (a range between 0-100) on indicators of treatment

(high- or low-deforestation), time period (prior to 1990 and 2000-2013), and relevant covariates (rainfall, temperature). Individual basin indicators or "fixed effects" account for time-invariant basin physical characteristics such as elevation and soil features. Because data are aggregated to period-month units, month fixed effects account for seasonally variable flow responses. Low flow percentiles represent flow that occurs in dry season months ($\sim$ July - September), middle-range flow percentiles represent flow that occurs in transition seasons ($\sim$ May - July and October - January), and high flow percentiles represent wet season flow ($\sim$ February - April). Based on data quality and results from a statistical "matching" analysis [150, 142] (see Methods and SI Text), which we used to evaluate the comparability of treatment and control basins, we evaluated the DID model on data from 46 river basins: a control group of 23 basins that experienced low levels of deforestation in the 2000-2013 period ($< 5\%$ of basin area deforested), and a treatment group of 23 basins that experienced relatively high levels of deforestation in the same period ($> 10\%$ of basin area deforested). All basins had less than 20% mean agricultural land cover prior to 1990, providing relative comparability with respect to baseline levels of deforestation.

The primary measure or regression coefficient of interest, which is called the "treatment effect" but is more formally known as the average treatment effect on the treated (ATT) [151], provides an estimate of average change in the flow of treated basins relative to control basins over the same time period. Results (Figure 4.3 and Table B.3) show that river basins with high-levels of deforestation between 2000-2013 had significantly greater change (increases) in low to mid-range rates of flow, relative to basins with low-levels of deforestation over the same period of time. For example, on average and holding all else (climate) constant, basins experiencing high-levels of deforestation experienced an increase in 5th percentile (dry season) flow that was 11 percentage points greater than change experienced by basins with low levels of deforestation.

We compliment the DID model with a similarly-specified linear mixed effects regression model [152], which is nearly identical to the DID model, but includes random intercepts for "nested" basin groups, and sites within nested groups, instead of basin fixed effects (see SI Text). Nested basin groups are collections of basins located along the same river network with overlapping drainage areas. The linear mixed effects model achieves results nearly equivalent to the DID model, albeit with different confidence intervals, across the entire range of mean-normalized flow percentiles (0-100th), showing a decline in the intensity of the effect with increasing rates of flow (Figure 4.3). Within the range of significant effects (0th-50th percentile of normalized flow), the treatment effect was between 0.05 and 0.11, meaning that increases in (mean-normalized) river flow in basins with $> 10\%$ of their area deforested were between 5 and 11 percentage points higher than in minimally deforested basins.

DID results agree with previous studies in individual basins in Brazil that identified the effect of transitions from natural to agricultural vegetation (pasture and soy) on low flow (baseflow) only [7, 8], a finding attributed to minimal soil infiltration effects (which would, if present, affect peak runoff or stormflow [6]) relative to significant evapotranspiration effects (which affect baseflow, particularly in dry periods). Consistency between these larger-sample

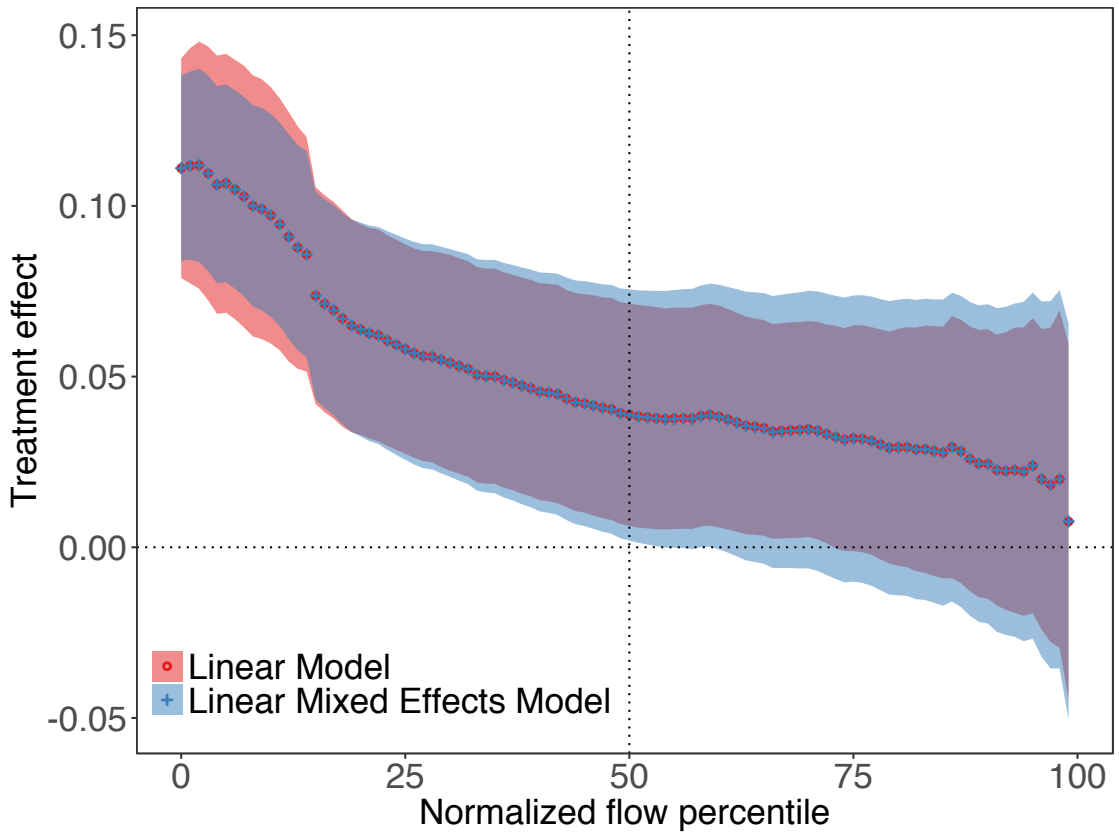

Figure 4.3: The estimated effects of deforestation on low to high rates of streamflow

The figure shows values of the average treatment effect (vertical axis) of high levels of deforestation (> 10% basin area deforested, 2000-2013) relative to low levels of deforestation (< 5% basin area deforested, 2000-2013) for 0-100th percentile mean-normalized flow (horizontal axis). The treatment effect was calculated using a fixed-effects ("Linear Model") and complimentary mixed effects ("Linear Mixed Effects Model") regressions, and outfit with data from 46 river basins over two periods: before 1990 and after 2000. The Linear Model represents a difference-in-differences (DID) approach; the mixed-effects model is nearly identical, but accounts for correlation between nested basins. The dotted vertical line marks the approximate flow percentile (50th) over which the average treatment effect is not significantly different than zero (according to a 95% confidence interval).

statistical findings and previous fieldwork suggest that site-specific findings of the dominance of evapotranspiration effects of deforestation are generalizable to the larger Amazon-Cerrado region. Nevertheless, significant stormflow effects may be present in intensively-grazed pasture and urbanized areas not well-represented in this study. Additionally, due to pre-existing agricultural land cover in some basins, the treatment effect is likely an under-estimate relative to what would be observed in only untouched basins undergoing deforestation (see SI Text).

Changes in annual streamflow

The previous analysis discretized deforestation in order to carry out a formal causal analysis, which provided information on the causal effects of deforestation relative to baseline observations across a range of flow rates. Alternately, continuous measurements allow a "dose-response" type quantification of streamflow rate responses to different levels of land use change. We use a linear mixed effect modeling approach [152] (see Methods and SI Text) to estimate the relationship between land cover (forest or agricultural, as a percent of basin area) and annual (water year, October - September) streamflow totals (percent of basin mean annual flow). This analysis uses two mixed effects regression models: the first, the "forest model", includes annual basin forest cover between 2000-2012 as a predictor; the second, the "agriculture model", includes annual agricultural land cover between 1950-2012 as a predictor. Both include basin physical and climate variables, and random effects for nested basin groups and sites within those groups.

Assuming our process understanding is correct (i.e. Figure 4.2), and holding all else constant (e.g. rainfall, streamflow, etc.), these empirical statistical models provide estimates of the effect of contemporary forest loss, and long-term agricultural land development - a proxy for historical deforestation, on conditional mean flow. The fitted models demonstrate that flow increases with forest cover loss and agricultural land cover gain (Table B.5 and Figure 4.4). If basin climate and physical features are fixed, then for each decrease of one percentage point in basin forest cover, annual streamflow (as a percent of basin mean annual flow) increase on average by 0.64 (0.80, 0.47) percentage points. Similarly, for each increase of one percentage point in basin agricultural land cover, conditional mean annual streamflow increases by 1.11 (0.44, 1.69) percentage points. According to the two models, an annual increase of 1cm of rainfall would generate a roughly similar magnitude of annual streamflow change, and corresponds on average to between a 0.57 (0.54, 0.60) to 0.75 (0.69, 0.80) percentage point increase in mean-normalized streamflow, according to agriculture and forest models, respectively. (Brackets indicate 95% confidence intervals.)

We estimate historical average flow change volumes (mm/year) using the fitted models and observed data, which allows for approximation of mean annual change for all basins and years in the time periods corresponding to forest cover (2000-2012) and agricultural land use (1950-2012), even though actual flow observations may be inconsistently distributed across years. This analysis did not require a continuous long-term record in any one basin, and is effectively a space-for-time substitution approach. We estimate average flow change - in

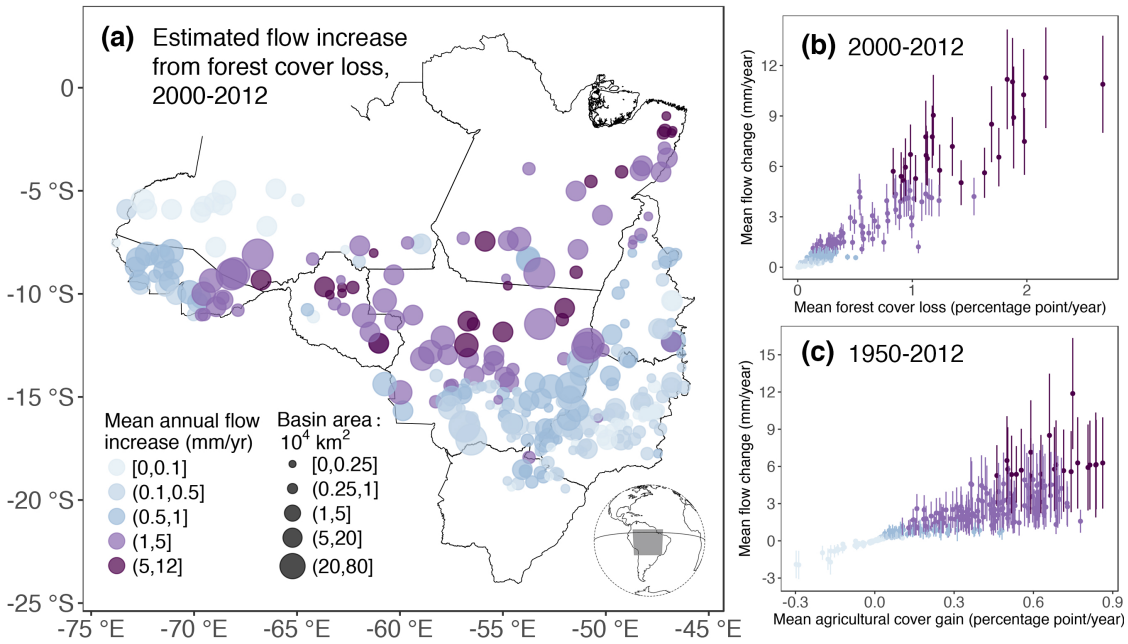

Figure 4.4: The estimated effects of deforestation and agricultural land development on annual flow

Panel (a) shows a map of (326) river basin centroids [32]; colors indicate annual flow increases (mean mm/year, 2000-2012) due to forest cover loss [132]; circle sizes indicate the size of the river basin. Panel (b) shows annual flow increases (mean mm/year, 2000-2012) due to forest cover loss plotted against mean annual forest cover loss by basin (corresponding to (a)). Similarly, panel (c) shows annual flow increases (mean mm/year, 1950-2012) due to agricultural land cover gain [131]. In (a) and (b) points are basin mean estimates, and error bars are 95% confidence intervals. Estimates are mean predicted values of change calculated from mixed effects model coefficient estimates (see Methods).

full and due to land cover change - in all basins using the fitted model (estimated regression coefficients) and continuous records of land cover, rainfall, temperature, as well as basin physical features (see Methods).

According to model estimates, forest loss between 2000-2012 (Figure 4.4 (a,b)) corresponds to a median flow increase across all 326 basins of 0.51 (0.37, 0.65) mm/year, which represents approximately 12% of the median of total estimated flow change (from all environmental drivers) of 4.18 mm/year (-2.04, 11.10). Agricultural cover gain between 1950-2012 (Figure 4.4 (c)) corresponds to a median flow increase across all 326 basins of 0.93 (0.37,1.57) mm/year, which represents approximately 44% of median total estimated flow change of 2.12 (-5.12,10.10) mm/year. (Numbers in brackets are bootstrap confidence intervals for flow change estimates, and bootstrap prediction intervals for flow change totals, the latter of which show substantial uncertainty).

Long term total flow change as estimated by the agricultural model is preferable due to its reliance on a longer record period. Nevertheless, the forest change model more directly

estimates the relationship between forest cover and flow, includes more recent and therefore higher quality hydroclimate data only, and is suggestive of a primary role of climate in generating flow change in recent years. The majority of flow change according to our estimates is due to rainfall variability, which includes any land cover change feedbacks to climate, which may be significant and may either reduce or amplify streamflow change [15, 16, 17, 20, 21]. In the agricultural model, the extent of agricultural frontier basins with substantial flow increases includes more southern basins (Figure B.9), which are those that were developed prior to the 2000-2012 period. Total flow change does not necessarily follow the same spatial pattern of land cover change-induced flow change, and can be found to decrease or increase in areas with even high rates of deforestation-altered flow (Figures B.11 and B.10)).

Naïve estimates of total flow change made directly from basin data are lower. According to a simple difference between average flow in 1965-1985 and 1992-2012 in 37 of the largest non-nested basins, the median rate of total flow change is 0.69 mm/year; according to simple linear regression of annual trends in 30 non-nested basins with more than 30 years of flow records, the median trend is 0.24 mm/year. These are similar to our mixed effects model estimates of flow change from land cover change *only*, but are nevertheless within the prediction interval provided for total flow change from the mixed effects model. Either sample limitations of actual basin flow records downwardly bias these naïve estimates of actual total flow change, or these results demonstrate a limitation of the empirical modeling method: it does not account for physical process limitations, i.e. annual basin water budgets, and therefore over-estimates change. In the latter case, results remain indicative of overall trends, and relative proportions of land cover-induced flow change to total flow change (as both would be similarly over-estimated), but total volume changes should be interpreted with caution.

Previous studies provide some insight on this matter. Our naïve estimates of total flow change correspond to data-based estimates of total flow change across South America between 1900-2000 (0.43 mm/year) [125]. Given the concentration of South America's land cover change in the Amazon-Cerrado region and in the last half century in particular, it is reasonable to expect rates of flow change to be higher on average in the region and time period of this study, supporting the higher rate estimates made using the mixed effects models. A land surface modeling exercise for the period of 1950-2000 estimated a change of 3,869 km^3/year in global runoff due to anthropogenic land cover change, relative to simulated potential natural vegetation coverage [153]. According to our agricultural model result applied over the total (non-overlapping basin) area of analysis (3.16 million km^2), we estimate a total volumetric rate of increase of 2.94 km^3/year between 1950-2012 in the Amazon-Cerrado region due to agricultural land cover gain (synonymous with forest and savanna woodland loss), which is 7.6% of the modeled total global rate of increase between 1950-2000 due to anthropogenic land cover change [153]. This is not unreasonable given the extent of land cover change in the region. Our estimated total volumetric rate of change, due to all environmental change - including land cover change, is 6.72 km^3/year between 1950-2012.

4.3 Conclusion

This study demonstrates that agricultural-driven deforestation in the Amazon-Cerrado region of Brazil is extensive in its impact to streamflow, particularly low, dry season flow. These findings are consistent with previous fieldwork- and model-based analysis in the region. A causal connection between forest loss and streamflow increase was demonstrated formally with a difference-in-differences (DID) analysis, showing that increases in (mean-normalized) river flow in basins with high levels of deforestation were between 5 and 11 percentage points higher than in minimally deforested basins, and providing a proof of concept for the use of observational (non-experimental) causal identification methods in the water sciences. Notably, larger-sample empirical methods allow for the estimation of gradual, less extreme land use change effects on streamflow where traditional methods are limited to analysis in more extreme land cover transition cases. An empirical mixed effects modeling exercise further quantified flow change due to land cover change and other environmental drivers, demonstrating significant annual rates of flow change across the arc of deforestation, and suggesting that on average between 1950-2012, the Amazon-Cerrado region experienced an increase of 0.93 mm/year (2.94 km^3/year) due to land cover change, accounting for 44% of total flow rate increases (6.72 km^3/year).

While elevated flow, and low flow in particular, may be seen as a desirable outcome of deforestation with respect to small, run-of-river hydropower generation and irrigation development in rural regions, tradeoffs between short-term benefits and long-term costs of (perceived) flow increases, including the masking of what would otherwise be decreases, are unclear. Deforestation has and will continue to slow in the region. Deforestation-driven flow increases may be sustained in the long term (such as in cultivated cropland), or flow may eventually decline due to regrowth (such as in partially-deforested natural pasture) [128]. What will happen in this region is unknown, but is likely a combination of the two outcomes. If this region experiences reduced rainfall and increased temperature, due to vegetation removal (reduced ET) or climate change [16, 9, 27], and flow increases taper with slowed deforestation, then compounded effects could result in a substantial and rapid decline in baseflow across the region in coming decades.

4.4 Methods

Data

We derived in-situ hydrological and climate data from the Brazilian Agência Nacional de Águas (ANA) and Agência Nacional de Energia Elétrica (ANEEL), which are provided in the custom "Curated Rain and Flow Data for the Brazilian Rainforest-Savanna Transition Zone" data package [32]. Land cover data include annual agricultural land use (including natural pasture, planted pasture, and annual and perennial cropland) [131] and forest cover and loss data [132]. In-situ rainfall data in Brazil [32], supplemented with samples from the

gridded global remotely-sensed rainfall product PERSIANN-CDR v.1 [89], was was inverse-distance weighting (IDW) interpolated at a 0.25 degree spatial resolution and daily temporal resolution between 1983-2013, and summarized to monthly and/or yearly resolution for analysis; rainfall in years prior to 1983 is from the gridded global rainfall product GPCC v.7 [154, 80]; custom interpolations of in-situ data are the preferred rainfall data source due to their superior performance in basin-level analyses [155]. Other data sources include: BEST daily temperature (C/day, mean and anomaly) [156]; SRTM v.4 elevation (m above sea level) [52]; and HWSD soil features including sand and clay fractions and organic carbon content (% weight) averaged over top- and sub-soils [157]. Data are area-weighted averages over river basin spatial units. (See SI Text and Table B.1 for details). 326 river basins met data quality requirements for analysis; additional criteria for inclusion in DID analyses further subset basins into a smaller set of 46 (SI Text).

Differences-in-differences (DID) analysis

Treated basins were those with more than 10% basin-area mean forest loss between 2000-2013, and control basins were those with less than 5% basin area-mean forest loss between 2000-2013; all treatment and control basins had $< 20\%$ agricultural land cover prior to 1990 in order to preserve correspondence between forest cover loss and agricultural land cover gain (Figure B.3), which is otherwise complicated by urbanization in basins with earlier development. The DID analysis relies on summaries of river basin longitudinal data over two time periods: before and after treatment. The post-treatment period is more precisely a period of gradual, continuous land use change that culminates in threshold levels of deforestation; flow and other variables are summarized over both periods, prior to and then simultaneous with this gradual change. Thus, there are 24 observations per river basin - 12 in each period for each month (data summarized over all available period-years in each month). Selected basins were required to have a minimum of five years of data for each month in the pre-treatment (< 1990) and post-treatment (2000-2013) periods. We used Mahalanobis Distance Matching (MDM) [158] to subset selected basins to those that were maximally balanced [142, 159, 160] with respect to time-invariant and/or pre-treatment qualities including: basin area, elevation, soil characteristics, rainfall, temperature, and agricultural land cover (see SI Text). The matching process resulted in selection of 46 basins for DID analysis. Specifications of the DID fixed effects log-linear regression, complimentary mixed effects linear regression model, and the treatment effect coefficient of interest, are described in the SI Text.

Streamflow Volumes

The mixed effects linear regression model used to estimate the relationship between mean-normalized annual total (depth) of streamflow and either forest or agricultural land cover is provided in the SI Text. River basins included in model fitting (173 basins for forest cover, and 91 basins for agricultural land cover) were limited to those with: at least ten

years of annual flow totals (across all available years) with fewer than three days missing in any given year; a minimum of five site-year observations; and inclusion in a nested group for which more than ten group-year observations were available in the measurement period (2000-2013 for forest cover and 1950-2012 for agricultural land cover). The agricultural land cover model was fit only on basins with less than 20% agricultural development by area prior to 1990; 1964 is the earliest record of flow in these selected basins. Predictions (estimated annual flow, and change in annual flow) were made on the full set of 326 basins for both forest (2000-2012) and agricultural land cover (1950-2012) data, which allows for evaluation of average flow change based on basin physical and climate features for all years, despite missing values in the observed flow record. For basins used in model fitting, predicted values rely on fixed effects and basin- and nested-group random effects; for basins not included in model fitting, predicted values use only fixed effects. Basin-specific mean annual flow values are used to back out volumes of flow from mean-normalized values. The proportion of flow change induced by land cover change relative to all environmental change is the ratio of flow change estimated from land cover change (regression coefficients) only, to flow change estimated with the full fitted model. See SI Text and Table B.5.

Software and code availability

All data formatted for this analysis, as well as code, are available upon request from the corresponding author. We carried out all analyses and generated all figures within the Comprehensive R Archive Network (CRAN) [55] programming environment (see SI Text for details).

Chapter 5

Conclusion

5.1 Summary of Findings

Chapter 1 introduces background information on land use and hydrology in Brazil's arc of deforestation across the Amazon rainforest and Cerrado (tropical savanna) biomes, and summarizes concepts and findings from Chapters 2 - 4.

Chapter 2 details effort required to acquire and process in-situ hydrological data, including rainfall, streamflow, and associated geographic information including river basin boundaries and the location and drainage areas of hydropower reservoirs. Curation of quality-controlled climate and hydrological data is an important but often overlooked 'data scientific' contribution to environmental research. Data science refers to the combination of traditional statistical data analysis, data processing (i.e. cleaning and formatting), and data visualization for communication. Data science thus acknowledges the conceptual and computational challenges posed by the structure, size, messiness, and complexity of data [34, 35], and the unique knowledge set required to address those challenges. Data science is not concerned with the overall conceptual framing of research or scientific methods, but the process by which data are made ready for analysis and the way in which results are interpreted. In order to test a hypothesis, a researcher must collect data, evaluate the quality of and clean data, and format data for a specific model or method. To present results of an analysis (e.g. a model or method output), the researcher must carry out robustness checks, quantify uncertainty, and visualize and summarize results for communication. In some studies, these tasks may be of adequate simplicity, or rely on datasets or methods with strong historical precedent, so as to be secondary in effort to the core scientific inquiry. In others, however, data challenges are paramount: the effort spent on what is traditionally known as the science may be small compared to that spent on the data science. The latter is the nature of the data scientific work of this dissertation, and is the reason for explicit presentation of methods used to curate the data used for analysis in Chapters 3 and 4.

Chapter 3 uses the rainforest-savanna transition region in Brazil as a case study to show differences in the statistics describing rainfall across nine remotely-sensed (RS) and

interpolated in-situ (IS) daily rainfall datasets covering the period of 1998-2013. Results from this study show that differences between rainfall datasets were comparable to estimated bias in global climate model projections, and rainfall trends from different datasets were inconsistent at the scale of river basins. We demonstrate that direct empirical comparisons between rainfall and streamflow provide a scalable alternative to modeling for evaluating rainfall dataset performance across multiple areal (river basin) units, and highlight the need for users of rainfall datasets to justify data choices for hydroclimatological analyses because analysts may reach divergent understandings due to data selection alone. Analysts can reduce 'data selection uncertainty' by (i) justifying dataset choices using selection methods such as the performance statistics demonstrated in Chapter 3, and/or (ii) including estimates of data-selection uncertainty (e.g. confidence intervals based on variability across datasets) in their findings. Evaluation of rainfall data prior to hydroclimatological analysis is both feasible (if streamflow records are available) and necessary. In contrast to the use of climate model outputs in analyses - where characterization of an ensemble of equally uncertain projections is best practice - if an individual dataset corresponds more closely with a reference of choice (e.g. streamflow) than other datasets, that dataset should be used for analysis.

Chapter 4 frames the case of land use change in Brazil as a natural experiment, and estimates the direct causal effect of deforestation on sreamflow within an observational data setting. This study demonstrates that agricultural-driven deforestation in the Amazon-Cerrado region of Brazil is extensive and regional in its impact to streamflow, particularly low, dry season flow. Results show that increases in (mean-normalized) river flow in basins with high levels of deforestation were between 5 and 11 percentage points higher than in minimally deforested basins, and that on average between 1950-2012, the Amazon-Cerrado region experienced an increase of 0.93 mm/year (2.94 km^3/year) due to land cover change, accounting for 44% of total flow rate increases (6.72 km^3/year). While elevated flow, and low flow in particular, may be seen as a desirable outcome of deforestation with respect to small, run-of-river hydropower generation and irrigation development in rural regions, tradeoffs between short-term benefits and long-term costs of (perceived) flow increases, including the potential masking of what would otherwise be decreases, are unclear. Deforestation-driven flow increases may be sustained in the long term (such as in cultivated cropland), or flow may eventually decline due to regrowth (such as in partially-deforested natural pasture) [128]. What will happen in this region is unknown, but is likely a combination of the two outcomes.

5.2 Future Work

The data outlined in Chapter 1 provides a foundation for a wide range of environmental analyses in rural and ecologically important Brazil. These core hydrological data can be combined with other spatially-explicit data products, such as elevation, soil maps, and remotely-sensed climate products (as in Chapters 3 and 4) to provide not only validation of remotely-sensed products themselves (see Chapter 3), but can also be employed in novel

empirical data analyses of hydroclimatic change (see Chapter 4) and in the calibration and validation of coupled land surface-atmosphere model outputs.

Future work stemming directly from this dissertation includes analyses of the effects of land use change on hydrology in several capacities additional to those presented in Chapter 4: (i) evaluation of basin scale effects of land use change on flow through the analysis of nested basin groups; (ii) evaluation of differences in flow (change) across basins with different types of agricultural land cover (e.g. pasture vs. double-cropped corn and soy) using additional land cover datasets providing detailed agricultural classifications [161, 94], and across basins with different spatial patterns (e.g. clustering) of land cover; (iii) evaluation of interactions of evapotranspiration (ET) and groundwater change in determining dynamics of flow change using additional remotely-sensed datasets, including MODIS ET [162, 163] and GRACE groundwater anomaly [164]; (iv) further investigation of the role of rainfall variability and temperature in determining flow change - those hinted at in the regression analyses of Chapter 4 that demonstrated significant effects of rainfall and temperature, especially with respect to climate and land use change feedbacks to climate; (v) the contribution of land cover change, via spatially variable evapotranspiration change effects, to observed temperature changes in the region; (vi) the relevance of flow change to regional hydropower generation, and especially small (run-of-river) hydropower; and (vii) modeling of future change in flow with (projected) future land cover change.

This study, and Chapter 4 in particular, provides a proof of concept for the use of observational (non-experimental) causal identification methods in the water sciences. Larger-sample empirical methods allow for the estimation of gradual, less extreme land use change effects on streamflow where traditional methods are limited to analysis in more extreme land cover transition cases. Thus, this dissertation motivates the discussion of the use of causal statistical analysis methods in the water sciences, and motivates a more complete discussion of the background and theory supporting causal statistical identification methods, their overlap with existing literature and empirical approaches in physical hydrology, and the translation of statistical terminology from other fields (economics, public health) to water science applications. This will also be the subject of future work.

Bibliography

[1] M. J. Lathuillire, M. T. Coe, and M. S. Johnson. "A review of green- and blue-water resources and their trade-offs for future agricultural production in the Amazon Basin: what could irrigated agriculture mean for Amazonia?" *Hydrol. Earth Syst. Sci.* 20.6 (June 7, 2016), pp. 2179–2194.

[2] Paulo Tarso S. Oliveira et al. "Trends in water balance components across the Brazilian Cerrado". *Water Resources Research* 50.9 (Sept. 1, 2014), pp. 7100–7114.

[3] Stephanie A. Spera et al. "Land-use change affects water recycling in Brazil's last agricultural frontier". *Global Change Biology* 22.10 (Oct. 1, 2016), pp. 3405–3413.

[4] M. T. Coe et al. "The effects of deforestation and climate variability on the streamflow of the Araguaia River, Brazil". *Biogeochemistry* 105.1 (Sept. 1, 2011), pp. 119–131.

[5] Lvia Cristina Pinto Dias et al. "Effects of land cover change on evapotranspiration and streamflow of small catchments in the Upper Xingu River Basin, Central Brazil". *Journal of Hydrology: Regional Studies* 4, Part B (Sept. 2015), pp. 108–122.

[6] Alphonce C. Guzha et al. "Characterizing rainfall-runoff signatures from micro-catchments with contrasting land cover characteristics in southern Amazonia". *Hydrological Processes* 29.4 (Feb. 15, 2015), pp. 508–521.

[7] Shelby J. Hayhoe et al. "Conversion to soy on the Amazonian agricultural frontier increases streamflow without affecting stormflow dynamics". *Global Change Biology* 17.5 (2011), pp. 1821–1833.

[8] Christopher Neill et al. "Watershed responses to Amazon soya bean cropland expansion and intensification". *Philosophical Transactions of the Royal Society of London B: Biological Sciences* 368.1619 (June 5, 2013), p. 20120425.

[9] Prajjwal K. Panday et al. "Deforestation offsets water balance changes due to climate variability in the Xingu River in eastern Amazonia". *Journal of Hydrology* 523 (Apr. 2015), pp. 822–829.

[10] Daniel C. Nepstad et al. "The role of deep roots in the hydrological and carbon cycles of Amazonian forests and pastures". *Nature* 372.6507 (Dec. 15, 1994), pp. 666–669.

[11] R. S. Oliveira et al. "Deep root function in soil water dynamics in cerrado savannas of central Brazil". *Functional Ecology* 19.4 (2005), pp. 574–581.

[12] L. S. Borma et al. "Atmosphere and hydrological controls of the evapotranspiration over a floodplain forest in the Bananal Island region, Amazonia". *Journal of Geophysical Research: Biogeosciences* 114 (G1 Mar. 1, 2009), G01003.

[13] Raphael Scheffler et al. "Soil hydraulic response to land-use change associated with the recent soybean expansion at the Amazon agricultural frontier". *Agriculture, Ecosystems & Environment* 144.1 (Nov. 2011), pp. 281–289.

[14] E. M. Latrubesse et al. "The geomorphologic response of a large pristine alluvial river to tremendous deforestation in the South American tropics: The case of the Araguaia River". *Geomorphology*. Short and Long Term Processes, Landforms and Responses in Large Rivers 113.3 (Dec. 15, 2009), pp. 239–252.

[15] Deborah Lawrence and Karen Vandecar. "Effects of tropical deforestation on climate and agriculture". *Nature Climate Change* 5.1 (Jan. 2015), pp. 27–36.

[16] Leticia S. Lima et al. "Feedbacks between deforestation, climate, and hydrology in the Southwestern Amazon: implications for the provision of ecosystem services". *Landscape Ecology* 29.2 (Feb. 2014). WOS:000331935100007, pp. 261–274.

[17] Alvaro Salazar et al. "Land use and land cover change impacts on the regional climate of non-Amazonian South America: A review". *Global and Planetary Change* 128 (May 2015), pp. 103–119.

[18] Leydimere J. C. Oliveira et al. "Large-scale expansion of agriculture in Amazonia may be a no-win scenario". *Environmental Research Letters* 8.2 (June 1, 2013), p. 024021.

[19] D. V. Spracklen, S. R. Arnold, and C. M. Taylor. "Observations of increased tropical rainfall preceded by air passage over forests". *Nature* 489.7415 (Sept. 13, 2012), pp. 282–285.

[20] D. V. Spracklen and L. Garcia-Carreras. "The impact of Amazonian deforestation on Amazon basin rainfall". *Geophysical Research Letters* 42.21 (Nov. 16, 2015), p. 2015GL066063.

[21] Abigail L. S. Swann et al. "Future deforestation in the Amazon and consequences for South American climate". *Agricultural and Forest Meteorology* 214215 (Dec. 15, 2015), pp. 12–24.

[22] Leandro Castello and Marcia N. Macedo. "Large-scale degradation of Amazonian freshwater ecosystems". *Global Change Biology* 22.3 (Mar. 1, 2016), pp. 990–1007.

[23] Carlos A. Nobre et al. "Land-use and climate change risks in the Amazon and the need of a novel sustainable development paradigm". *Proceedings of the National Academy of Sciences* 113.39 (Sept. 27, 2016), pp. 10759–10768.

[24] Jacson Hudson Incio Ferreira et al. "Assessment of the potential of small hydropower development in Brazil". *Renewable and Sustainable Energy Reviews* 56 (Apr. 2016), pp. 380–387.

[25] Alexander C. Lees et al. "Hydropower and the future of Amazonian biodiversity". *Biodiversity and Conservation* 25.3 (Mar. 9, 2016), pp. 451–466.

[26] Fernando Almeida Prado Jr. et al. "How much is enough? An integrated examination of energy security, economic growth and climate change related to hydropower expansion in Brazil". *Renewable and Sustainable Energy Reviews* 53 (Jan. 2016), pp. 1132–1136.

[27] Claudia M. Stickler et al. "Dependence of hydropower energy generation on forests in the Amazon Basin at local and regional scales". *Proceedings of the National Academy of Sciences* (May 13, 2013), p. 201215331.

[28] World Wide Inland Navigation Network. *Brazil Inland Waterways*. http://www.wwinn.org/brazil-inland-waterways. 2016.

[29] K. O. Winemiller et al. "Balancing hydropower and biodiversity in the Amazon, Congo, and Mekong". *Science* 351.6269 (Jan. 8, 2016), pp. 128–129.

[30] Fabio Aprile and Assad Jos Darwich. "Nutrients and water-forest interactions in an Amazon floodplain lake: an ecological approach". *Acta Limnologica Brasiliensia* 25.2 (June 2013), pp. 169–182.

[31] Morgan C. Levy. *Curated rain and flow data for the Brazilian rainforest-savanna transition zone*. https://dx.doi.org/10.6084/m9.figshare.3100912. Figshare, 2016.

[32] Morgan C. Levy. *Curated rain and flow data for the Brazilian rainforest-savanna transition zone*. http://www.hydroshare.org/resource/e82e66572b444fc5b6bf16f88f911f77. Consortium of Universities for the Advancement of Hydrologic Science, Hydroshare, 2016.

[33] Christopher Hutton et al. "Most computational hydrology is not reproducible, so is it really science?" *Water Resources Research* 52.10 (Oct. 1, 2016), pp. 7548–7555.

[34] Vasant Dhar. "Data science and prediction". *Communications of the ACM* 56.12 (Dec. 1, 2013), pp. 64–73.

[35] Rachel Schutt and Cathy O'Neil. *Doing data science*. 2013.

[36] Agência Nacional de Águas (ANA). *Sistema Nacional de Informações sobre Recursos Hdricos (SNIRH). Rede Hidrometeorológica Nacional*. http://www.ana.gov.br/PortalSuporte/frmSelecaoEstacao.aspx. 2016.

[37] Agência Nacional de Águas (ANA). *Geospatial Metadata Portal. GeoNetwork open-source*. http://metadados.ana.gov.br/geonetwork/srv/pt/main.home. Nov. 6, 2013.

[38] Agência Nacional de Águas (ANA), Superintendence of Management of the Hydro meteorological Network. *Orientations for consistency of rainfall data*. Tech. rep. http://arquivos.ana.gov.br/infohidrologicas/cadastro/OrientacoesParaConsistenciaDadosPluviometricos-VersaoJul12.pdf. Brasilia, Brazil: Agência Nacional de Águas (ANA), 2012, p. 21.

[39] Departamento Nacional de Águas e Energia Elétrica, Divisão de Controle de Recursos Hídricos DCRH. *Sistemtica para Anlise de Consistncia e Homogeneizao de Dados Pluviomtricos*. Tech. rep. Braslia, DF, Brazil: Departamento Nacional de Águas e Energia Elétrica (DNAEE), 1984.

[40] World Meteorological Organization. *Guide to Meteorological Instruments and Methods of Observation*. Tech. rep. WMO - No. 8. http://library.wmo.int/pmb_ged/wmo_8_en-2012.pdf. Geneva, Switzerland: World Meteorological Organization, 2008, p. 716.

[41] Jean-Marie Bouroche and Gilbert Saporta. *L'analyse des donnes*. Presses Universitaires de France - PUF, Nov. 2010.

[42] Brian Everitt. *Cluster analysis*. Reviews of current research: 11. London : Heinemann Educational [for] the Social Science Research Council, 1974.

[43] Joe H. Ward Jr. "Hierarchical Grouping to Optimize an Objective Function". *Journal of the American Statistical Association* 58.301 (Mar. 1963), pp. 236–244.

[44] J.C. Bertoni and C.E.M. Tucci. "Precipitation". *Hydrology: Science and Application*. Ed. by Carlos Eduardo Morelli Tucci. 2nd ed. Porto Alegre, Brazil, 2001.

[45] James K. Searcy, Clayton H. Hardison, and Walter B. Langein. *Double-mass curves; with a section fitting curves to cyclic data*. USGS Numbered Series 1541-B. U.S. Govt. Print. Off., 1960.

[46] Gérard Hiez. "L'homogénéité des données pluviométriques". *Cahiers ORSTOM. Srie Hydrologie* 14.2 (1977), pp. 129–172.

[47] G. Hiez. *Processamento dos dados pluviométricos do nordeste: homogeneização dos dados métodos do vetor regional*. Tech. rep. Recife, Brazil: Superintendência de Desenvolvimento do Nordeste, 1978.

[48] Agência Nacional de Águas (ANA), Superintendence of Management of the Hydro meteorological Network. *Orientations for consistency of streamflow data*. Tech. rep. http://arquivos.ana.gov.br/infohidrologicas/cadastro/OrientacoesParaConsistenciaDadosFluviometricos-VersaoJul12.pdf. Brasilia, Brazil: Agência Nacional de Águas (ANA), 2012, p. 19.

[49] Departamento Nacional de Águas e Energia Elétrica, Divisão de Controle de Recursos Hídricos DCRH. *Sistemtica para Anlise de Consistncia de Dados Fluviomtricos*. Tech. rep. Braslia, DF, Brazil: Departamento Nacional de Águas e Energia Elétrica (DNAEE), 1982.

[50] S.E. Rantz. *Measurement and Computation of Streamflow. Volume 2. Computation of Discharge*. Tech. rep. USGS Water Supply Paper 2175. USGS, 1982.

[51] DHV Consultants BV and Delft Hydraulics. *How to establish stage discharge rating curve*. Tech. rep. Training module # SWDP - 29. New Delhi, India, Nov. 1999, p. 31.

[52] A. Jarvis et al. *Hole-filled SRTM for the globe Version 4, available from the CGIAR-CSI SRTM 90m Database*. http://www.cgiar-csi.org/data/srtm-90m-digital-elevation-database-v4-1. 2008.

[53] Agência Nacional de Energia Elétrica (ANEEL). *Sistema de Informações Georreferenciadas do Setor Elétrico (SIGEL)*. http://sigel.aneel.gov.br/sigel.html. 2016.

[54] Sistema de Informações Georreferenciadas do Setor Elétrico (SIGEL). Agência Nacional de Energia Elétrica (ANEEL). *Download KMZ - SIGEL - ANEEL*. http://sigel.aneel.gov.br/kmz.html. 2016.

[55] R Core Team. *R: A Language and Environment for Statistical Computing*. https://www.R-project.org. Vienna, Austria: R Foundation for Statistical Computing, 2016.

[56] Vadlamudi Brahmananda Rao, Iracema F. A. Cavalcanti, and Kioshi Hada. "Annual variation of rainfall over Brazil and water vapor characteristics over South America". *Journal of Geophysical Research: Atmospheres* 101.D21 (Nov. 1996), pp. 26539–26551.

[57] V. Brahmananda Rao et al. "An update on the rainfall characteristics of Brazil: seasonal variations and trends in 19792011". *International Journal of Climatology* 36.1 (Jan. 1, 2016), pp. 291–302.

[58] P. C. D. Milly et al. "Stationarity Is Dead: Whither Water Management?" *Science* 319.5863 (Feb. 2008), pp. 573–574.

[59] C. J. Vörösmarty et al. "Global threats to human water security and river biodiversity". *Nature* 467.7315 (Sept. 2010), pp. 555–561.

[60] Elizabeth E. Ebert, John E. Janowiak, and Chris Kidd. "Comparison of Near-Real-Time Precipitation Estimates from Satellite Observations and Numerical Models". *Bulletin of the American Meteorological Society* 88.1 (Jan. 1, 2007), pp. 47–64.

[61] Maria Gehne et al. "Comparison of Global Precipitation Estimates across a Range of Temporal and Spatial Scales". *Journal of Climate* 29.21 (July 29, 2016), pp. 7773–7795.

[62] Viviana Maggioni, Patrick C Meyers, and Monique D Robinson. "A Review of Merged High-Resolution Satellite Precipitation Product Accuracy during the Tropical Rainfall Measuring Mission (TRMM) Era". *Journal of Hydrometeorology* 17.4 (2016), pp. 1101–1117.

[63] Evangelia-Anna Kalognomou et al. "A diagnostic evaluation of precipitation in CORDEX models over southern Africa". *Journal of Climate* 26.23 (2013), pp. 9477–9506.

[64] D.A. Hughes and A. Slaughter. "Daily disaggregation of simulated monthly flows using different rainfall datasets in southern Africa". *Journal of Hydrology: Regional Studies* 4, Part B (2015), pp. 153 –171.

[65] Aurore Degre, Sarann Ly, and Catherine Charles. "Different methods for spatial interpolation of rainfall data for operational hydrology and hydrological modeling at watershed scale: a review". *Biotechnology, Agronomy, Society and Environment* 17.2 (Aug. 2013), pp. 392–406.

[66] Marc Girons Lopez et al. "Location and Density of Rain Gauges for the Estimation of Spatial Varying Precipitation". *Geografiska Annaler: Series A, Physical Geography* 97.1 (Mar. 2015), pp. 167–179.

[67] Stephanie K. Kampf and Stephen J. Burges. "Quantifying the water balance in a planar hillslope plot: Effects of measurement errors on flow prediction". *Journal of Hydrology* 380.1–2 (Jan. 2010), pp. 191–202.

[68] Lisa C. Sieck, Stephen J. Burges, and Matthias Steiner. "Challenges in obtaining reliable measurements of point rainfall". *Water Resources Research* 43.1 (Jan. 1, 2007), 10.1029/2005WR004519.

[69] P. Karimi and W. G. M. Bastiaanssen. "Spatial evapotranspiration, rainfall and land use data in water accounting Part 1: Review of the accuracy of the remote sensing data". *Hydrology and Earth System Sciences* 19.1 (Jan. 28, 2015), pp. 507–532.

[70] Alan Mair and Ali Fares. "Comparison of Rainfall Interpolation Methods in a Mountainous Region of a Tropical Island". *Journal of Hydrologic Engineering* 16.4 (2011), pp. 371–383.

[71] Fengge Su, Yang Hong, and Dennis P. Lettenmaier. "Evaluation of TRMM Multisatellite Precipitation Analysis (TMPA) and Its Utility in Hydrologic Prediction in the La Plata Basin". *Journal of Hydrometeorology* 9.4 (Aug. 2008), pp. 622–640.

[72] Franois Massonnet et al. "Using climate models to estimate the quality of global observational data sets". *Science* 354.6311 (Oct. 28, 2016), pp. 452–455.

[73] R Core Team. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing, 2015.

[74] David M. Olson et al. "Terrestrial Ecoregions of the World: A New Map of Life on Earth". *BioScience* 51.11 (Nov. 2001), pp. 933–938.

[75] Letícia S. Lima et al. "Feedbacks between deforestation, climate, and hydrology in the Southwestern Amazon: implications for the provision of ecosystem services". *Landscape Ecology* 29.2 (2014), pp. 261–274.

[76] Marta Llopart et al. "Climate change impact on precipitation for the Amazon and La Plata basins". *Climatic Change* 125.1 (July 2014), pp. 111–125.

[77] Claudia M. Stickler et al. "Dependence of hydropower energy generation on forests in the Amazon Basin at local and regional scales". *Proceedings of the National Academy of Sciences* 110.23 (2013), pp. 9601–9606.

[78] Oliver L. Phillips et al. "Drought Sensitivity of the Amazon Rainforest". *Science* 323.5919 (Mar. 2009), pp. 1344–1347.

[79] Ke Zhang et al. "The fate of Amazonian ecosystems over the coming century arising from changes in climate, atmospheric CO2, and land use". *Global Change Biology* 21.7 (July 2015), pp. 2569–2587.

[80] Udo Schneider et al. "GPCC's new land surface precipitation climatology based on quality-controlled in situ data and its role in quantifying the global water cycle". *Theoretical and Applied Climatology* 115.1 (Mar. 20, 2013), pp. 15–40.

[81] G Botter et al. "Basin-scale soil moisture dynamics and the probabilistic characterization of carrier hydrologic flows: Slow, leaching-prone components of the hydrologic response". *Water resources research* 43.2 (2007).

[82] G. Thirel et al. "Hydrology under change: an evaluation protocol to investigate how hydrological models deal with changing catchments". *Hydrological Sciences Journal* 60.7 (Aug. 3, 2015), pp. 1184–1199.

[83] Günter Blöschl. *Runoff prediction in ungauged basins: synthesis across processes, places and scales*. Cambridge University Press, 2013.

[84] Marc F. Muller and Sally E. Thompson. "Bias adjustment of satellite rainfall data through stochastic modeling: Methods development and application to Nepal". *Advances in Water Resources* 60 (Oct. 2013), pp. 121–134.

[85] Murugesu Sivapalan and Gnter Blschl. "Transformation of point rainfall to areal rainfall: Intensity-duration-frequency curves". *Journal of Hydrology* 204.1 (1998), pp. 150 –167.

[86] Bruno Collischonn, Walter Collischonn, and Carlos Eduardo Morelli Tucci. "Daily hydrological modeling in the Amazon basin using TRMM rainfall estimates". *Journal of Hydrology* 360.1–4 (Oct. 2008), pp. 207–216.

[87] A.S. Gebregiorgis and F. Hossain. "Understanding the Dependence of Satellite Rainfall Uncertainty on Topography and Climate for Hydrologic Model Simulation". *IEEE Transactions on Geoscience and Remote Sensing* 51.1 (Jan. 2013), pp. 704–718.

[88] Peter A. Troch et al. "Climate and vegetation water use efficiency at catchment scales". *Hydrological Processes* 23.16 (July 2009), pp. 2409–2414.

[89] Hamed Ashouri et al. "PERSIANN-CDR: Daily Precipitation Climate Data Record from Multisatellite Observations for Hydrological and Climate Studies". *Bulletin of the American Meteorological Society* 96.1 (June 27, 2014), pp. 69–83.

[90] Carla Gulizia and Inés Camilloni. "Comparative analysis of the ability of a set of CMIP3 and CMIP5 global climate models to represent precipitation in South America". *International Journal of Climatology* 35.4 (Mar. 2015), pp. 583–595.

[91] Vivek K Arora. "The use of the aridity index to assess climate change effect on annual runoff". *Journal of Hydrology* 265.1 (Aug. 30, 2002), pp. 164–177.

[92] Christopher L. Shope and Ganga Ram Maharjan. "Modeling Spatiotemporal Precipitation: Effects of Density, Interpolation, and Land Use Distribution". *Advances in Meteorology* 2015 (Apr. 2015), p. 174196.

[93] H. Otieno et al. "Influence of Rain Gauge Density on Interpolation Method Selection". *Journal of Hydrologic Engineering* 19.11 (2014), p. 04014024.

[94] Avery S. Cohn et al. "Cropping frequency and area response to climate variability can exceed yield response". *Nature Climate Change* 6.6 (June 2016), pp. 601–604.

[95] Mara Paula Llano and Walter Vargas. "Climate characteristics and their relationship with soybean and maize yields in Argentina, Brazil and the United States". *International Journal of Climatology* 36.3 (Mar. 1, 2016), pp. 1471–1483.

[96] Joseph L. Awange, Freddie Mpelasoka, and Rodrigo M. Goncalves. "When every drop counts: Analysis of Droughts in Brazil for the 1901-2013 period". *Science of The Total Environment* 566567 (Oct. 1, 2016), pp. 1472–1488.

[97] Keith Beven and Ida Westerberg. "On red herrings and real herrings: disinformation and information in hydrological inference". *Hydrological Processes* 25.10 (May 2011), pp. 1676–1680.

[98] Guillaume Thirel, Vazken Andréassian, and Charles Perrin. "On the need to test hydrological models under changing conditions". *Hydrological Sciences Journal* 60.7-8 (Aug. 2015), pp. 1165–1173.

[99] T. Cohen Liechti et al. "Comparison and evaluation of satellite derived precipitation products for hydrological modeling of the Zambezi River Basin". *Hydrology and Earth System Sciences* 16.2 (Feb. 2012), pp. 489–500.

[100] M. L. M. Scheel et al. "Evaluation of TRMM Multi-satellite Precipitation Analysis (TMPA) performance in the Central Andes region and its dependency on spatial and temporal resolution". *Hydrology and Earth System Sciences* 15.8 (Aug. 2011), pp. 2649–2663.

[101] Maximilian Auffhammer et al. "Using Weather Data and Climate Model Output in Economic Analyses of Climate Change". *Review of Environmental Economics and Policy* 7.2 (July 2013), pp. 181–198.

[102] Salvatore Pascale et al. "Analysis of rainfall seasonality from observations and climate models". *Climate Dynamics* 44.11-12 (Aug. 2014), pp. 3281–3301.

[103] Jessica D. Lundquist et al. "Diagnosis of insidious data disasters". *Water Resources Research* 51.5 (May 2015), pp. 3815–3827.

[104] George J. Huffman et al. "Global Precipitation at One-Degree Daily Resolution from Multisatellite Observations". *Journal of Hydrometeorology* 2.1 (Feb. 2001), pp. 36–50.

[105] Pingping Xie et al. "A Gauge-Based Analysis of Daily Precipitation over East Asia". *Journal of Hydrometeorology* 8.3 (June 2007), pp. 607–626.

[106] Mingyue Chen et al. "Assessing objective techniques for gauge-based analyses of global daily precipitation". *Journal of Geophysical Research: Atmospheres* 113.D4 (Feb. 2008), p. D04110.

[107] George J. Huffman et al. "The TRMM Multisatellite Precipitation Analysis (TMPA): Quasi-Global, Multiyear, Combined-Sensor Precipitation Estimates at Fine Scales". *Journal of Hydrometeorology* 8.1 (Feb. 1, 2007), pp. 38–55.

[108] Soroosh Sorooshian et al. "Evaluation of PERSIANN System Satellite–Based Estimates of Tropical Rainfall". *Bulletin of the American Meteorological Society* 81.9 (Sept. 2000), pp. 2035–2046.

[109] Columbia University, Earth Institute. *IRI/LDEO Climate Data Library*. 2015.

[110] Gopi Goteti. *raincpc: Obtain and Analyze Rainfall Data from the Climate Prediction Center*. R package version 0.4. 2014.

[111] Alfred H. Thiessen. "Precipitation averages for large areas". *Monthly Weather Review* 39.7 (July 1911), pp. 1082–1089.

[112] Donald Shepard. "A Two-dimensional Interpolation Function for Irregularly-spaced Data". *Proceedings of the 1968 23rd ACM National Conference*. ACM '68. New York, NY, USA: ACM, 1968, pp. 517–524.

[113] G Matheron. *Le krigeage universel*. [Paris, France]: École nationale supérieure des mines de Paris, 1969.

[114] Noel A. C Cressie. *Statistics for spatial data*. New York: Wiley, 1993.

[115] Robert M. Hirsch, James R. Slack, and Richard A. Smith. "Techniques of trend analysis for monthly water quality data". *Water Resources Research* 18.1 (Feb. 1, 1982), pp. 107–121.

[116] Robert M. Hirsch and James R. Slack. "A Nonparametric Trend Test for Seasonal Data With Serial Dependence". *Water Resources Research* 20.6 (June 1, 1984), pp. 727–732.

[117] Henry B. Mann. "Nonparametric Tests Against Trend". *Econometrica* 13.3 (July 1, 1945), pp. 245–259.

[118] Pranab Kumar Sen. "Estimates of the Regression Coefficient Based on Kendall's Tau". *Journal of the American Statistical Association* 63.324 (Dec. 1, 1968), pp. 1379–1389.

[119] Henri Theil. "A rank-invariant method of linear and polynomial regression analysis". *Henri Theils Contributions to Economics and Econometrics*. Springer, 1992, pp. 345–381.

[120] V. Lyne and M. Hollick. "Stochastic time variable rainfall-runoff modelling". *Proceedings of the Hydrology and Water Resources Symposium*. Perth, Australia: Institution of Engineers National Conference Publication, No. 79/10, pp. 89-92, Sept. 1979.

[121] Frédéric Ibanez. "Sur une nouvelle application de la théorie de 1'information à la description des séries chronologiques planctoniques". *Journal of Plankton Research* 4.3 (Jan. 1982), pp. 619–632.

[122] Maurice G Kendall. *Time-series*. New York: Hafner Press, 1976.

[123] Maurice G Kendall. *Rank correlation methods*. London: Griffin, 1970.

[124] C. Spearman. "The Proof and Measurement of Association between Two Things". *The American Journal of Psychology* 15.1 (Jan. 1904), pp. 72–101.

[125] Shilong Piao et al. "Changes in climate and land use have a larger direct impact than rising CO2 on global river runoff trends". *Proceedings of the National Academy of Sciences* 104.39 (Sept. 25, 2007), pp. 15242–15247.

[126] FAO. *Global Forest Resources Assessment 2015.* 2nd edition. Rome, Italy, 2016.

[127] INPE. *Projeto PRODES, Monitoramente da floresta amazonica brasileira por satelite*. So Paulo, Brazil, 2016.

[128] Alice E. Brown et al. "A review of paired catchment studies for determining changes in water yield resulting from alterations in vegetation". *Journal of Hydrology* 310.1 (Aug. 1, 2005), pp. 28–61.

[129] Michael T. Coe, Marcos H. Costa, and Britaldo S. Soares-Filho. "The influence of historical and potential future deforestation on the stream flow of the Amazon River Land surface processes and atmospheric feedbacks". *Journal of Hydrology* 369.1 (May 5, 2009), pp. 165–174.

[130] Marcos Heil Costa, Aurlie Botta, and Jeffrey A Cardille. "Effects of large-scale changes in land cover on the discharge of the Tocantins River, Southeastern Amazonia". *Journal of Hydrology* 283.1 (Dec. 10, 2003), pp. 206–217.

[131] Lvia C. P. Dias et al. "Patterns of land use, extensification, and intensification of Brazilian agriculture". *Global Change Biology* 22.8 (Aug. 1, 2016), pp. 2887–2903.

[132] M. C. Hansen et al. "High-Resolution Global Maps of 21st-Century Forest Cover Change". *Science* 342.6160 (Nov. 15, 2013), pp. 850–853.

[133] Yadvinder Malhi et al. "Climate Change, Deforestation, and the Fate of the Amazon". *Science* 319.5860 (Jan. 11, 2008), pp. 169–172.

[134] Ramdane Alkama and Alessandro Cescatti. "Biophysical climate impacts of recent changes in global forest cover". *Science* 351.6273 (Feb. 5, 2016), pp. 600–604.

[135] Christiane Runyan and Paolo D'Odorico. *Global deforestation*. New York, NY : Cambridge University Press, 2016., 2016.

[136] Qiuhong Tang and Taikan Oki. *Terrestrial water cycle and climate change : Natural and human-induced impacts*. Geophysical monograph series: 221. Washington, DC : American Geophysical Union, 2016, 2016.

[137] Vazken Andrassian. "Waters and forests: from historical controversy to scientific debate". *Journal of Hydrology* 291.1 (May 31, 2004), pp. 1–27.

[138] L. A. Bruijnzeel. "Hydrological functions of tropical forests: not seeing the soil for the trees?" *Agriculture, Ecosystems & Environment*. Environmental Services and Land Use Change: Bridging the Gap between Policy and Research in Southeast Asia 104.1 (Sept. 2004), pp. 185–228.

[139] Nicolas Zgre et al. "In lieu of the paired catchment approach: Hydrologic model change detection at the catchment scale". *Water Resources Research* 46.11 (2010), n/a–n/a.

[140] Armando Brath, Alberto Montanari, and Greta Moretti. "Assessing the effect on flood frequency of land use change via hydrological simulation (with uncertainty)". *Journal of Hydrology* 324.1 (June 15, 2006), pp. 141–153.

[141] Paul R Rosenbaum. *Design of observational studies*. OCLC: 663096465. New York: Springer, 2010.

[142] Guido Imbens and Donald B. Rubin. *Causal inference for statistics, social, and biomedical sciences : an introduction*. New York, NY, USA : Cambridge University Press., 2015.

[143] A. A. Chariton et al. "Emergent technologies and analytical approaches for understanding the effects of multiple stressors in aquatic environments". *Marine and Freshwater Research* 67.4 (2016), pp. 414–428.

[144] Ray Hilborn. "Correlation and Causation in Fisheries and Watershed Management". *Fisheries* 41.1 (Jan. 2, 2016), pp. 18–25.

[145] Caitlin E. Pearson et al. "Resolving large-scale pressures on species and ecosystems: propensity modelling identifies agricultural effects on streams". *Journal of Applied Ecology* 53.2 (Apr. 1, 2016), pp. 408–417.

[146] Song S. Qian and R. Daren Harmel. "Applying Statistical Causal Analyses to Agricultural Conservation: A Case Study Examining P Loss Impacts". *JAWRA Journal of the American Water Resources Association* 52.1 (Feb. 1, 2016), pp. 198–208.

[147] H. V. Gupta et al. "Large-sample hydrology: a need to balance depth with breadth". *Hydrol. Earth Syst. Sci.* 18.2 (Feb. 6, 2014), pp. 463–477.

[148] Egbert H. van Nes et al. "Causal feedbacks in climate change". *Nature Climate Change* 5.5 (May 2015), pp. 445–448.

[149] Joshua David Angrist and Jrn-Steffen Pischke. *Mostly harmless econometrics : an empiricist's companion*. Princeton : Princeton University Press, c2009., 2009.

[150] Donald B. Rubin. "Matching to Remove Bias in Observational Studies". *Biometrics* 29.1 (1973), pp. 159–183.

[151] Susan Athey and Guido W. Imbens. "Identification and Inference in Nonlinear Difference -in-Differences Models". *Econometrica* 74.2 (Mar. 1, 2006), pp. 431–497.

[152] Andrew Gelman and Jennifer Hill. *Data analysis using regression and multilevel/ hierarchical models.* Analytical methods for social research. Cambridge ; New York : Cambridge University Press, 2007., 2007.

[153] Shannon M. Sterling, Agns Ducharne, and Jan Polcher. "The impact of global land-cover change on the terrestrial water cycle". *Nature Climate Change* 3.4 (Apr. 2013), pp. 385–390.

[154] A. Becker et al. "A description of the global land-surface precipitation data products of the Global Precipitation Climatology Centre with sample applications including centennial (trend) analysis from 1901present". *Earth System Science Data* 5.1 (2013), pp. 71–99.

[155] M.C. Levy et al. "Addressing rainfall data selection uncertainty using connections between rainfall and streamflow". *Nature Scientific Reports* (In Review).

[156] Robert Rohde et al. "Berkeley Earth Temperature Averaging Process". *Geoinformatics & Geostatistics: An Overview* 01.2 (2013).

[157] FAO/IIASA/ISRIC/ISSCAS/JRC. *Harmonized World Soil Database (version 1.2).* FAO, Rome, Italy and IIASA, Laxenburg, Austria, 2012.

[158] Donald B. Rubin. "Using Multivariate Matched Sampling and Regression Adjustment to Control Bias in Observational Studies". *Journal of the American Statistical Association* 74.366 (1979), pp. 318–328.

[159] Stephen L. Morgan and Christopher Winship. *Counterfactuals and causal inference : methods and principles for social research.* Analytical methods for social research. New York, NY : Cambridge University Press, 2015., 2015.

[160] Elizabeth A. Stuart. "Matching methods for causal inference: A review and a look forward". *Statistical science : a review journal of the Institute of Mathematical Statistics* 25.1 (Feb. 1, 2010), pp. 1–21.

[161] Stephanie A. Spera et al. "Recent cropping frequency, expansion, and abandonment in Mato Grosso, Brazil had selective land characteristics". *Environmental Research Letters* 9.6 (May 1, 2014), p. 064010.

[162] Qiaozhen Mu et al. "Development of a global evapotranspiration algorithm based on MODIS and global meteorology data". *Remote Sensing of Environment* 111.4 (Dec. 28, 2007), pp. 519–536.

[163] Qiaozhen Mu, Maosheng Zhao, and Steven W. Running. "Improvements to a MODIS global terrestrial evapotranspiration algorithm". *Remote Sensing of Environment* 115.8 (Aug. 15, 2011), pp. 1781–1800.

[164] NASA Jet Propulsion Laboratory. *GRACE Tellus Gravity Recovery and Climate Experiment*. GRACE Tellus: Get Data. http://grace.jpl.nasa.gov/data/get-data. 2016.

[165] Hongfen Teng et al. "Estimating spatially downscaled rainfall by regression kriging using TRMM precipitation and elevation in Zhejiang Province, southeast China". *International Journal of Remote Sensing* 35.22 (2014), pp. 7775–7794.

[166] S. S. Shapiro and M. B. Wilk. "An Analysis of Variance Test for Normality (Complete Samples)". *Biometrika* 52.3 (1965), pp. 591–611.

[167] Tom Fawcett. "An Introduction to ROC Analysis". *Pattern Recogn. Lett.* 27.8 (June 2006), pp. 861–874.

[168] F. T. Andrews, B. F. W. Croke, and A. J. Jakeman. "An open software environment for hydrological model assessment and development". *Environmental Modelling & Software* 26.10 (Oct. 2011), pp. 1171–1185.

[169] Claude Elwood Shannon and Warren Weaver. *The mathematical theory of communication.* Urbana: University of Illinois Press, 1949.

[170] T. M Cover and Joy A Thomas. *Elements of information theory*. New York: Wiley, 1991.

[171] Roger S. Bivand, Edzer Pebesma, and Virgilio Gomez-Rubio. *Applied spatial data analysis with R, Second edition*. Springer, NY, 2013.

[172] Edzer J. Pebesma and Roger S. Bivand. "Classes and methods for spatial data in R". *R News* 5.2 (Nov. 2005), pp. 9–13.

[173] Robert J. Hijmans. *raster: Geographic Data Analysis and Modeling*. R package version 2.4-18. 2015.

[174] Edzer J. Pebesma. "Multivariable geostatistics in S: the gstat package". *Computers & Geosciences* 30 (2004), pp. 683–691.

[175] Nick Bond. *hydrostats: Hydrologic indices for daily time series data*. R package version 0.2.3. 2014.

[176] Philippe Grosjean and Frederic Ibanez. *pastecs: Package for Analysis of Space-Time Ecological Series*. R package version 1.3-18. 2014.

[177] Max Kuhn Contributions from Jed Wing et al. *caret: Classification and Regression Training*. R package version 6.0-52. 2015.

[178] Aldo Marchetto. *rkt: Mann-Kendall Test, Seasonal and Regional Kendall Tests*. R package version 1.4. 2015.

[179] Hadley Wickham. *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York, 2009.

[180] David Kahle and Hadley Wickham. "ggmap: Spatial Visualization with ggplot2". *The R Journal* 5.1 (2013), pp. 144–161.

[181] Adrian Baddeley and Thomas Lawrence. *globe: Plot 2D and 3D Views of the Earth, Including Major Coastline*. R package version 1.1-2. 2016.

[182] A. Carla Staver, Sally Archibald, and Simon A. Levin. "The Global Extent and Determinants of Savanna and Forest as Alternative Biome States". *Science* 334.6053 (2011-10-14), pp. 230–232.

[183] Gary King and Richard Nielsen. "Why Propensity Scores Should Not Be Used for Matching". *Working Paper* (2016).

[184] Robert J. Hijmans. *geosphere: Spherical Trigonometry*. R package version 1.5-5. 2016.

[185] Hadley Wickham. "Reshaping Data with the reshape Package". *Journal of Statistical Software* 21.12 (2007), pp. 1–20.

[186] Hadley Wickham and Romain Francois. *dplyr: A Grammar of Data Manipulation*. R package version 0.5.0. 2016.

[187] Douglas Bates et al. "Fitting Linear Mixed-Effects Models Using lme4". *Journal of Statistical Software* 67.1 (2015), pp. 1–48.

[188] Alexandra Kuznetsova, Per Bruun Brockhoff, and Rune Haubo Bojesen Christensen. *lmerTest: Tests in Linear Mixed Effects Models*. R package version 2.0-32. 2016.

[189] Jared E. Knowles and Carl Frederick. *merTools: Tools for Analyzing Mixed Effect Regression Models*. R package version 0.2.1. 2016.

Appendix A

Supplementary Information (SI): Chapter 3

A.1 Figures

All figures were generated using the Comprehensive R Archive Network (CRAN) [73] programming environment (Version 3) on both Apple and Windows operating systems. See the Supplementary Discussion for a list of utilized software packages. See the Methods section of the main text for data source references.

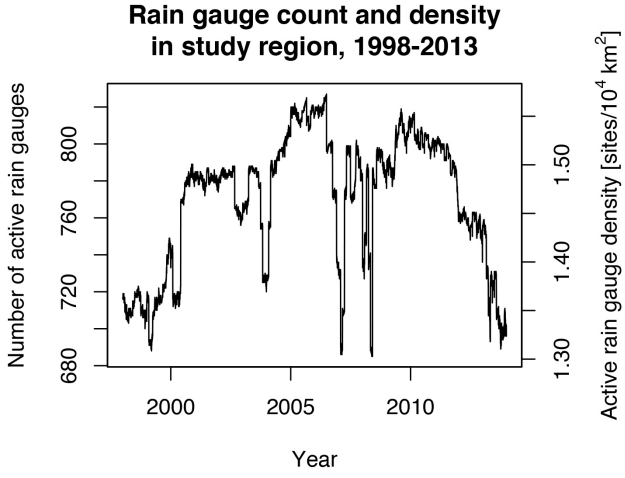

Figure A.1: Rain gauge counts and densities in study region

The number of active rain gauges (left vertical axis) and active rain gauge density (right vertical axis), daily between 1998-2013, in the eight-state and 5.25 million km^2 study region. Rain gauge data is from the curated data package used in this study [32].

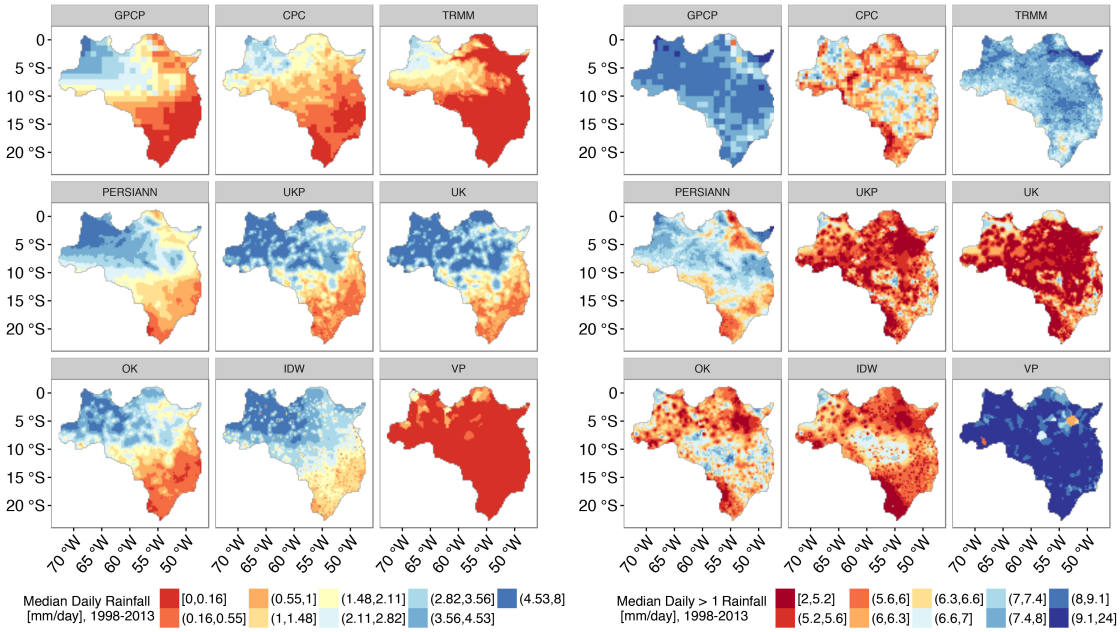

Figure A.2: Median and wet-day median rainfall

Median daily and wet-day (days with $\geq$ 1 mm/day) median daily rainfall between 1998-2013 over the study region according to different datasets. Median values were calculated at each 0.25° grid cell. These maps were generated in R, Version 3 (https://cran.r-project.org/) [55].

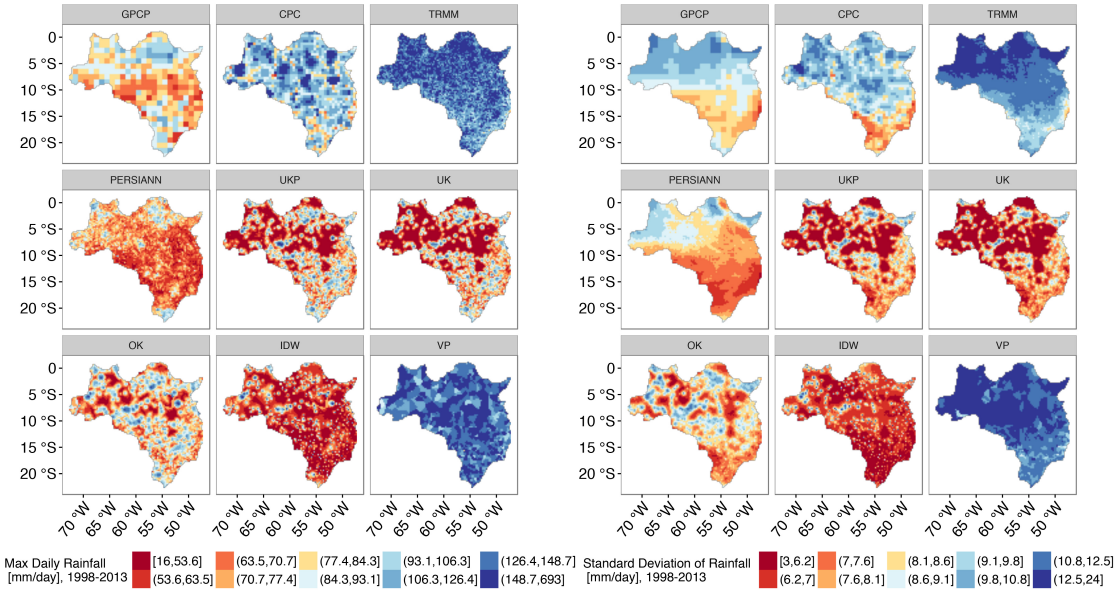

Figure A.3: Extremes and variability of rainfall

Maximum (left) and standard deviation (right) of daily rainfall between 1998-2013 over the study region according to different datasets. Maxima and standard deviations were calculated at each 0.25° grid cell. These maps were generated in R, Version 3 (https://cran.r-project.org/) [73].

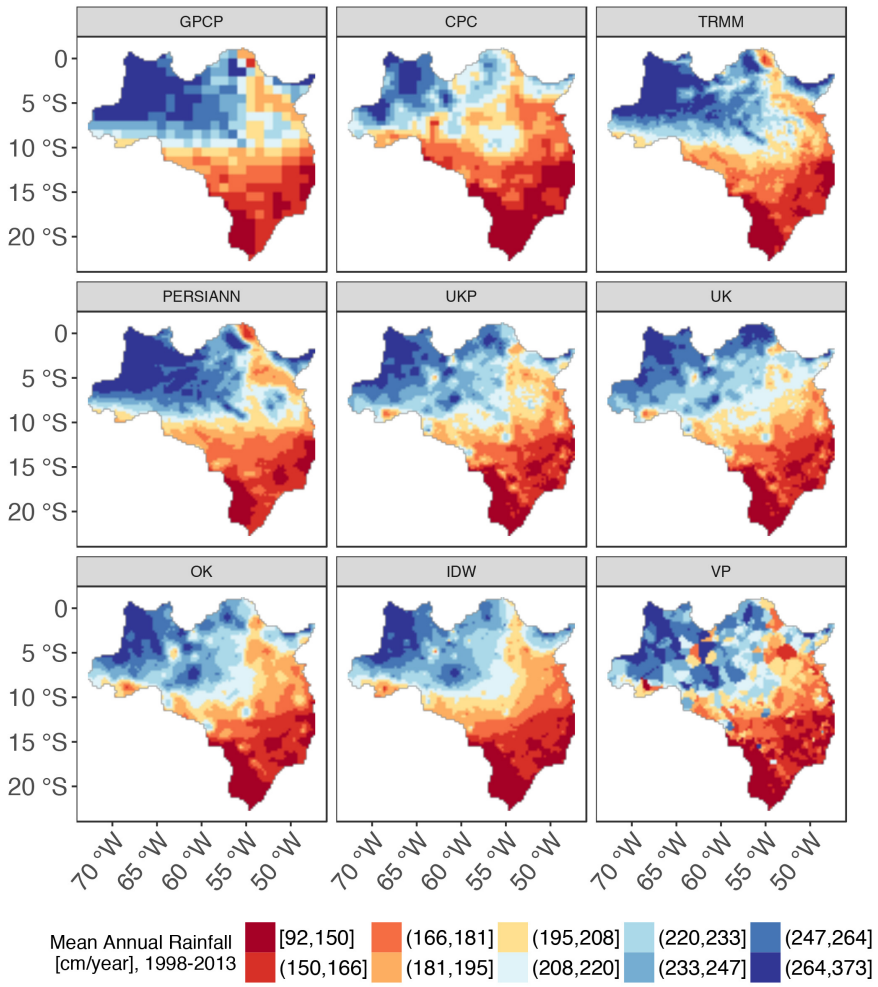

Figure A.4: Mean annual total rainfall

Mean annual water year (October-September) totals of rainfall between 1998-2013 over the study region according to different datasets. Mean values were calculated at each 0.25° grid cell. These maps were generated in R, Version 3 (https://cran.r-project.org/) [73].

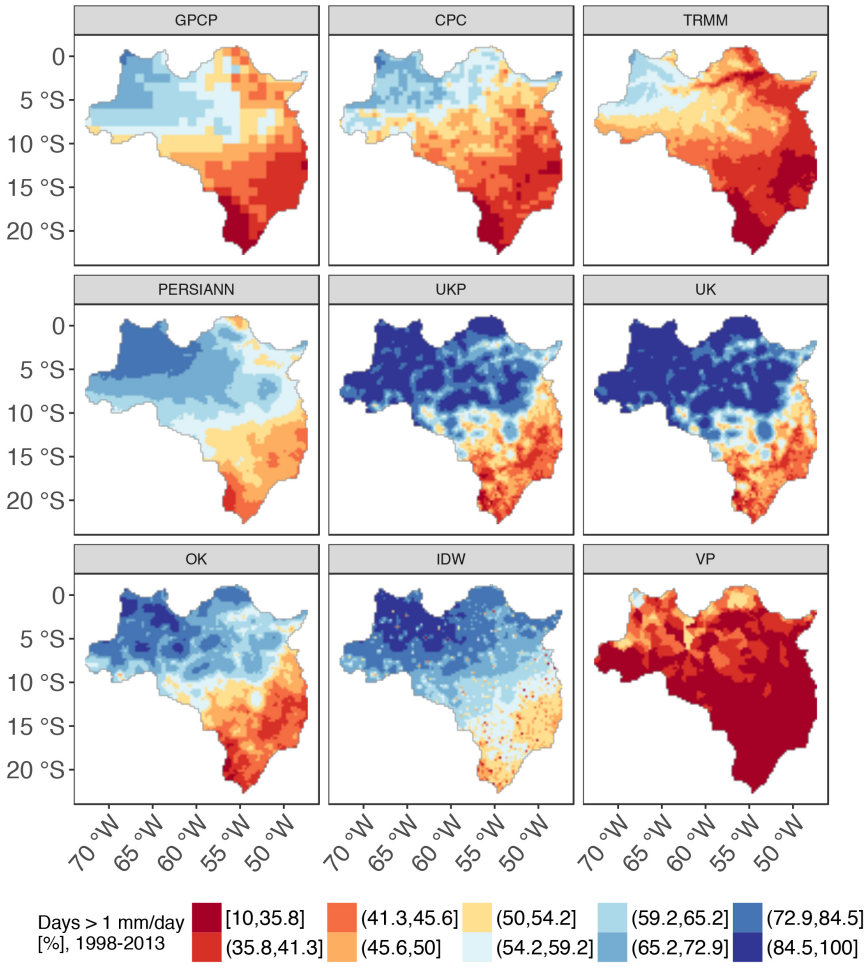

Figure A.5: Occurrence of rainfall

Daily rainfall occurrence (percent of wet days, when rainfall depths are $\geq$ 1mm) between 1998-2013 over the study region according to different datasets. Occurrence was calculated at each 0.25° grid cell. These maps were generated in R, Version 3 (https://cran.r-project.org/) [73].

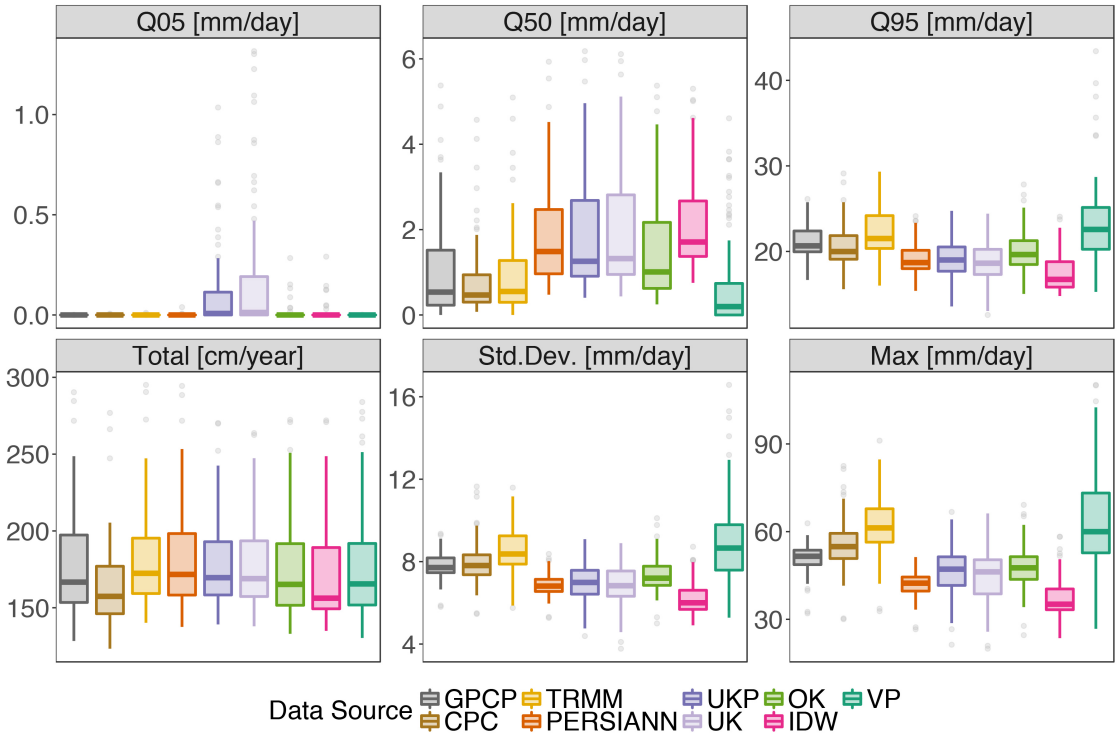

Figure A.6: Supplemental daily rainfall statistics in river basins according to different rainfall datasets

From left to right, top to bottom, the panels show the 5th, 50th (median), and 95th percentiles of daily rainfall (mm/day); the (mean) total annual rainfall (cm/year); standard deviation of daily rainfall (mm/day); and the (mean annual) maximum daily rainfall (mm/day). Each boxplot is generated with n=89 (river basin) statistic results, calculated using basin area-average rainfall from the given rainfall dataset (colors) from all days between 1998-2013.

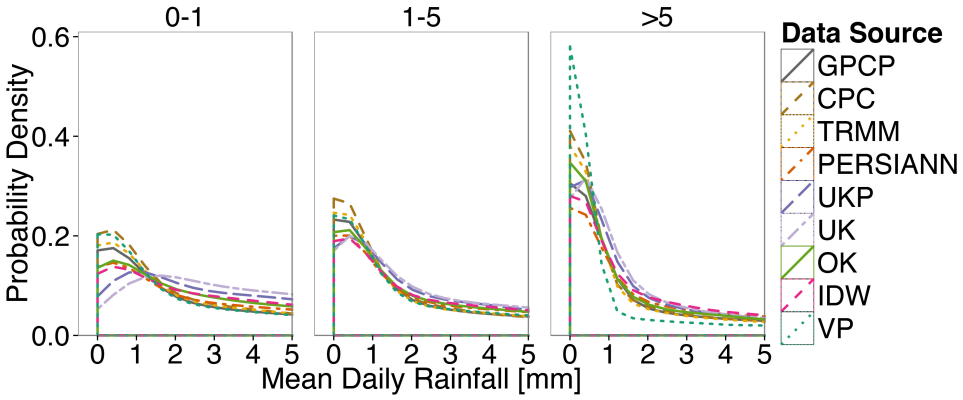

Figure A.7: Daily rainfall distributions by rain gauge density

Empirical probability distributions of daily rainfall between 1998-2013 across sample areas in the study region by rainfall dataset (curves) and rain gauge density bins (panels). Rain gauge density bins are labeled on the top of panels, and refer to the number of rain gauges per 10^4 km^2. Samples at different gauge densities are from rainfall-averaging areas extending outward from 100 regularly-sampled points in the study region with radii between 10-200 km. The number of observations in each gauge density bin was balanced by sub-sampling due to a greater number of observations in the low-density bin: each final bin was composed of 10^6 observations. Distributions deviate between datasets primarily at the lowest and highest rain gauge densities. GPCP, CPC, TRMM, and VP report a greater number of dry-day, low-rainfall depth events (< 1 mm/day) than UKP, UK, OK, IDW, and PERSIANN, especially at lower and higher gauge densities. The latter group of datasets register wet-day (≥ 1 mm/day), medium intensity rainfall depths more frequently than the first group, especially at the lowest and highest rain gauge densities.

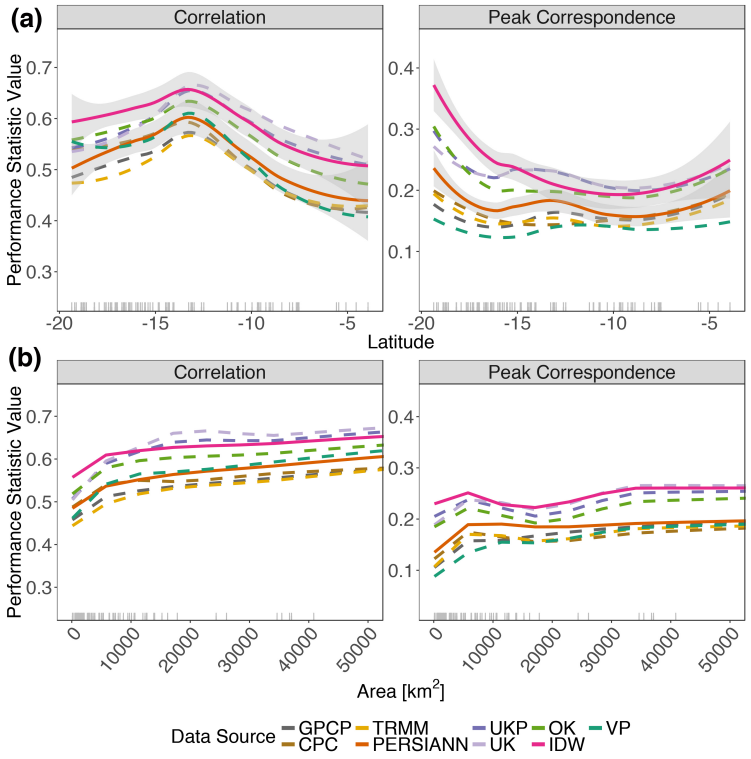

Figure A.8: Differences in rainfall data quality as indicated by performance statistics, by latitude and river basin area

Panel (a) shows performance statistics (correlation and peak correspondence) by rainfall dataset (colors) plotted as local regression-smoothed curves across the range of latitudes (at basin centroids) in the study region, with 95% uncertainty intervals (shaded). Panel (b) shows the same performance statistic curves plotted across the range of river basin area sizes in the study region. In both panels, solid lines indicate the best-performing gridded (PERSIANN) and custom-interpolated (IDW) datasets. In (a), non-overlapping uncertainty intervals indicate distinguishable performance between datasets; dashed-lines indicate all other datasets with uncertainty intervals that are not displayed, but are of similar width. In (b), uncertainty intervals overlap in most of the displayed range due to the sparse sampling of basin size across the range of basin sizes, and are therefore not shown. Gray tick-marks at the bottom illustrate the spread of latitudes and rain gauge densities in the 89 river basins.

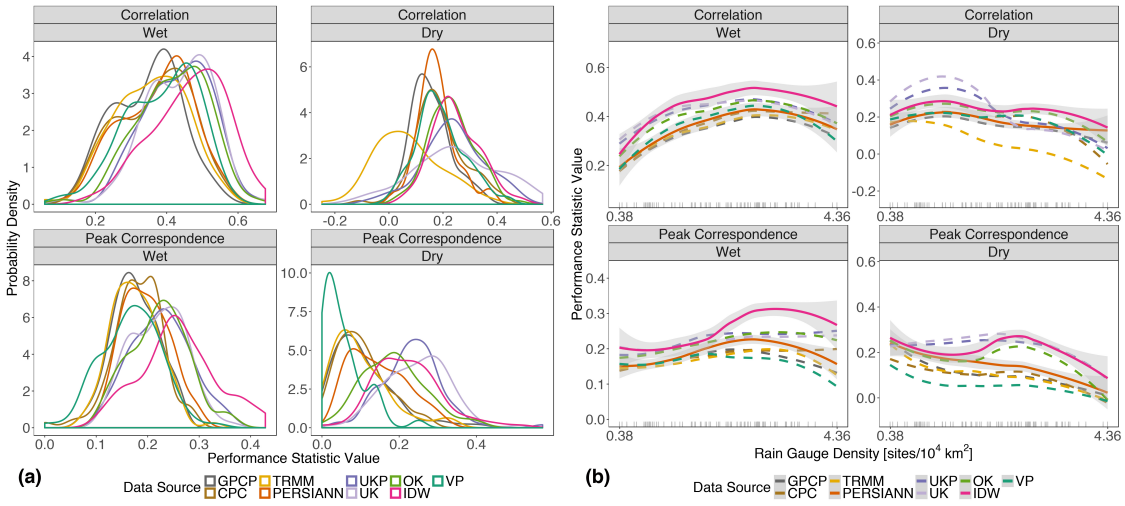

Figure A.9: Differences in rainfall data quality as indicated by performance statistics, by season

Panel (a) shows kernel-smoothed empirical probability distributions of performance statistics (correlation and peak correspondence) by rainfall dataset and season (wet: October - April, dry: May - September). Panel (b) shows the same performance statistics plotted by season as local regression-smoothed curves across the range of rain gauge densities in the study region, with 95% uncertainty intervals (shaded). In (b), solid lines and shaded regions indicate the best-performing gridded (PERSIANN) and custom-interpolated (IDW) datasets and their 95% uncertainty intervals, respectively; non-overlapping uncertainty intervals indicate distinguishable performance between datasets; dashed-lines indicate all other datasets with uncertainty intervals that are not displayed, but are of similar width; gray tick-marks at the bottom illustrate the spread of rain gauge densities in the 89 river basins.

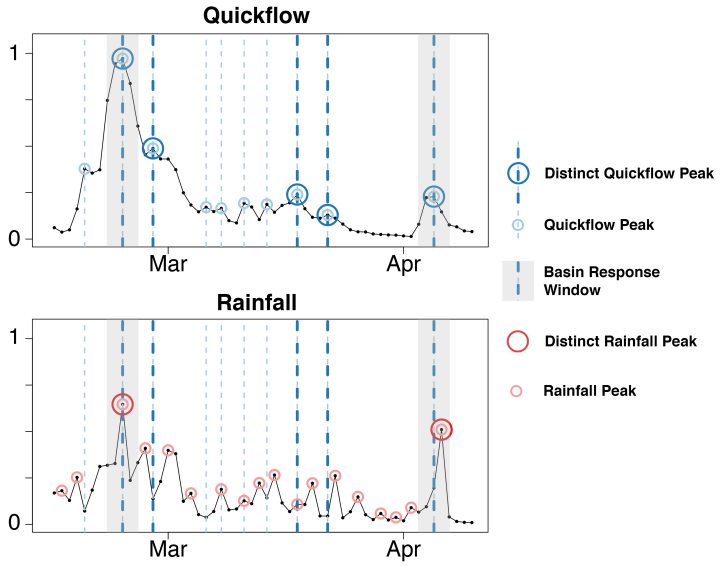

Figure A.10: Schematic of rainfall and streamflow peak correspondence methodology

Daily quickflow (top) and rainfall (bottom) for a two-month period; the data have been standardized to [0,1]; rainfall has been lagged to maximize rain and flow cross-correlation. 'Peak correspondence' is a performance statistic that measures the rate at which distinct rainfall peaks correspond to distinct quickflow peaks within a basin-specific response time window. The shaded windows illustrate cases where distinct peaks correspond; on the left - an exact match, and on the right - a match within the window.

A.2 Tables

Table A.1: Two-sample Kolmogorov-Smirnov tests for differences in distributions of performance statistics

Statistic	**Data**	*GPCP*	*CPC*	*TRMM*	*PERS*	*UKP*	*UK*	*OK*	*IDW*
Correlation	*CPC*	**0.034**							
	TRMM	0.3	**0.008**						
	PERS	0.078	0.988	**0.008**					
	UKP	**0**	**0.003**	**0**	**0.003**				
	UK	**0**	**0.001**	**0**	**0.001**	0.948			
	OK	**0**	**0.003**	**0**	**0.001**	0.868	0.222		
	IDW	**0**	**0**	**0**	**0**	0.113	0.16	**0.034**	
	VP	**0.014**	0.868	**0.001**	0.948	**0.005**	**0.002**	**0.008**	**0**
Peak Corresp.	*CPC*	0.3							
	TRMM	0.756	0.3						
	PERS	**0.003**	0.16	**0.003**					
	UKP	**0**	**0**	**0**	**0**				
	UK	**0**	**0**	**0**	**0**	0.988			
	OK	**0**	**0**	**0**	**0**	0.113	0.078		
	IDW	**0**	**0**	**0**	**0**	0.16	0.3	**0.005**	
	VP	**0.022**	**0.003**	0.052	**0**	**0**	**0**	**0**	**0**

Two-sample Kolmogorov-Smirnov test p-values (rounded to three decimal places) for differences in distributions of performance statistics (for year-round data). Datasets with performance statistic distributions that are significantly different from each other are bolded (p-value < 0.05).

A.3 Supplementary discussion

Additional details about rainfall data

Rain gauge density factors into the quality of gridded datasets to the extent that each dataset relies on gauge data: RS datasets (GPCP, TRMM, PERSIANN) rely on gauge data for calibration, and the gridded IS dataset (CPC) relies on gauge data entirely. There is overlap in the gauge data used in custom interpolations (from the custom data package [32]) and quality-controlled data released by the Brazilian government and included in the National Center for Atmospheric Research (NCAR) Global Precipitation Climatology Centre (GPCC) monthly product, which is used to calibrate all three RS products evaluated in this study. The exact number of overlapping gauges across all gridded and custom interpolated products is not available as data creators do not release detailed source data. Based on information provided in referenced source documents (see Methods), our custom IS interpolations incorporate gauged locations that are additional to those used in the gridded datasets, and incorporate higher temporal resolution gauge data directly (instead of inclusion of only monthly IS data for calibration, such as in the RS products [61]). Therefore it is possible that the quality of custom IS interpolations relative to gridded datasets (according to performance statistics) derives primarily from the greater number of rain gauges used in the IS interpolations. Confirmation of this assertion for gridded RS datasets (GPCP, TRMM, PERSIANN) would require an ability to separate the contribution of satellite retrievals (and

associated processing algorithms) from the contribution of rain gauge calibration to the overall product; for the gridded IS dataset (CPC), confirmation would require use of the CPC interpolation method on the custom data. These tasks are beyond the scope of this study, and therefore we do not answer this question.

Our custom interpolations of IS data provide estimates of total rainfall amounts that are similar to gridded data (see Supplementary Figures A.4 and A.6), especially when aggregated up to monthly time scales. No single dataset reports consistently higher or lower rainfall depths at any given location. According to statistics calculated over large to small sample areas (100 regularly-sampled locations across the study region with radii of 200 km and 10 km, respectively): IDW, OK, and UKP report the lowest proportion of extreme rainfall (highest and lowest value) and anomalous rainfall (relative to mean of all datasets combined) at daily, monthly, and annual time scales. All gridded datasets and VP, and in some cases UK, are responsible for a greater fraction of extreme and anomalous rainfall estimates (although no single dataset is responsible for $> 50\%$). The datasets providing the most extreme/anomalous estimates (and the percentage of extremes/anomalies they are responsible for) change with respective to the temporal resolution (daily, monthly, annual). According to cumulative sum plots, seasonality is consistent across the different datasets.

Selection of IS interpolation resolution

In an analysis of rain gauge (pair) distances across the study region between 1998-2013, the average median distance is 1,150 km; the average 5% quantile distance is 250 m, and the average 95% quantile distance is 2,370 km. While the average minimum distance is 2km, this distance occurs for only a handful of rain gauge locations. Across the whole study region, most areas have rain gauges located at a distance of greater than 1,000 km, for which a 0.25° (28 km) resolution is sufficient. In a preliminary analysis, we carried out local IS interpolations over basin areas at a 0.05° resolution; our analysis with those interpolations yielded equivalent results, thus the performance of IS interpolations relative to RS data is not a function of interpolation resolution.

Selection of IS interpolation specifications

We selected IDW parameters, UK predictor variables (latitude, longitude, elevation [52], and RS data - PERSIANN), and compared prediction error of the different interpolation methods and UK specifications using k-fold cross validation (CV). CV results were evaluated using correlation, error interquartile ranges, and RMSE of the errors normalized by standard deviation of observations. Kruskal-Wallis tests confirmed that these metrics discriminated between the interpolation methods' performance. An IDW parameter of 1.5, and UK (UKP) covariates of latitude, longitude, and elevation (and PERSIANN) were best performing. The UKP interpolation we used combines RS and IS data, which has been recommended [100, 165].

Additional information on hydroclimate indices

Normality tests (q-q plots and the Shapiro-Wilk test [166]) suggest that the runoff ratio and Horton index values (for each rainfall dataset) and/or means (across rainfall datasets) may be non-normal, demonstrating that alternative methods for the calculation of confidence intervals (e.g. non-parametric bootstrap sampling) may be preferred.

Additional information on performance statistics

IDW performs best on average, but the universal kriging methods (UKP and UK) can outperform IDW at low gauge densities. Unlike IDW and OK, which are local interpolations, UKP and UK generate mean rainfall values at locations with no nearby rain gauges (using 'universal' predictors such as elevation across the full study region), thereby increasing the occurrence of nonzero rainfall values at those locations. Because a greater number of nonzero rainfall values will correspond with flow rises (relevant to correlation) and peaks (relevant to peak correspondence) more frequently, kriging methods demonstrate better performance at very low gauge densities. Therefore, this result does not necessarily imply better representation of rainfall by kriging methods in low gauge density areas. Lastly, the kriging method incorporating RS data (UKP) performs no better than its IS-only counterpart (UK), although inclusion of the RS data in UKP attenuates UK's otherwise higher median daily values and occurrence.

With respect to the peak correspondence response time windows: in windows exceeding $1/3 \times \tau$ (τ is the basin response timescale in units of days), the peak correspondence method fails to distinguish between datasets. Use of a static window of 1 and 2 days across all basins yielded similar results to those presented in the main text (where a basin-specific window of $1/4 \times \tau$ was used). In the context of a contingency or error matrix analysis [167], peak correspondence is the true-positive rate (TPR) at which distinct peaks match within a window of time that accounts for potential mismatches in the estimated basin response time. Thus, 1-TPR gives a 'false-negative' rate, corresponding to type-II error, or error of omission, in the rainfall datasets.

The performance statistics were validated on an external set of rainfall and streamflow data from seven Australian basins of various sizes, including those with yearlong (wet) and intermittent (dry) flow of different periods of record (2 - 39 years), provided in the R `hydromad` package [168]. Real rainfall from each basin was perturbed with normally-distributed additive random noise (mean = 0, standard deviation equal to a range of between 0.25 to 2 times the rainfall standard deviation) to create 100 synthetic rainfall datasets per basin; negative values of perturbed rainfall were set to zero, thus synthetic rainfall included additional rainfall events, as well as eliminated existing ones. The generation of 100 synthetic datasets was repeated 100 times (iterations) for each of the Australian basins. For each of the 100 iterations for each basin, we calculated the percentage of times each performance statistic (correlation and peak correspondence) correctly identified the real rainfall dataset from the combined set of synthetic (100) and real (1) datasets. Correlation and peak correspondence

reliably selected the real rainfall dataset across basins with different flow types (intermittent and yearlong) and time series lengths; correlation was, however, better able to distinguish the correct dataset when random perturbations were small (low signal to noise ratios). When the standard deviation was set to 1-2 times the real rainfall standard deviation, both performance statistics correctly identified the real rainfall dataset 100% of the time. When the standard deviation was smaller, for example, 0.5 times the rainfall standard deviation, correlation identified the correct rainfall dataset again 100% of the time, however peak correspondence identified the correct rainfall dataset on average (over all basins) 79% of the time. Specifically, peak correlation identified the correct rainfall dataset in two of the seven basins 100% of the time, regardless of the signal to noise ratio. However, it identified the correct rainfall dataset with decreasing accuracy in other basins as the signal to noise ratio was decreased. There were no similar features (e.g. size, intermittency) in the basins for which peak correspondence worked best or worst (although short record durations negatively affect peak correspondence due its reliance on peak events, which may occur infrequently, especially in dry basins). Thus, we assume that variation in the ability of peak correspondence to identify the correct rainfall dataset has to do with the nature of runoff data from individual basins; basins in which peak correspondence is most likely to work well, given low signal to noise ratios, are those that produce clear (and ideally, frequent) quickflow response signals.

We also evaluated a suite of alternative performance statistics, including: mutual information [169, 170]; peak correspondence calculated with less distinct (higher probability) peaks (e.g. using all, instead of just distinct peaks); normalized cumulative distances; peak correspondence and correlation with respect to local regression-smoothed rainfall time series; peak correspondence and correlation with respect to probability- and information-weighted quickflow and rainfall peaks; and moving-window correlation of peak magnitudes for rainfall and quickflow. Mutual information results were similar to correlation, but were less reliable in identifying true rainfall when tested on the suite of external data (described above), as were peak correspondence measures calculated with less distinct peaks. The remaining trial statistics either failed to reliably identify the correct rainfall dataset in validation tests, or duplicated the results of the two primary performance statistics.

Software

Formatting and analysis of spatial data relied on the core R spatial analysis packages `sp` [171, 172], `raster` [173], `gstat` [174] and their dependencies. To estimate delay times between rainfall and flow events and obtain test data for performance statistic validation, we used the `hydromad` package [168]. To perform baseflow and quickflow separation, we used the `hydrostats` package [175]. To categorize peaks in rainfall and quickflow, and calculate their probability and significance, we used the `pastecs` package [176]. For construction and analysis of error matrices, we used the `caret` package [177]. We used the `rkt` package [178] for Seasonal Kendall (SK) tests and trend estimators. Figures and maps were generated using using the `ggplot2` [179], `ggmap` [180], and `globe` [181] packages and their dependencies.

Appendix B

Supplementary Information (SI): Chapter 4

B.1 Text

Data

Table B.1 lists all data, its temporal and spatial resolution, time period coverage, and source. In-situ hydrological data is provided along with comprehensive documentation on data acquisition and quality assurance/quality control in the "Curated Rain and Flow Data for the Brazilian Rainforest-Savanna Transition Zone" data package [32] [see private link], and an analysis-ready subset of these data are made discoverable through the Consortium of Universities for the Advancement of Hydrologic Science (CUAHSI) Hydroshare platform (https://www.cuahsi.org/HydroShare).

326 river basins met basic data quality requirements, which included: having high quality streamflow data (see documentation in [32]); being an unregulated river basin, which was defined as not having a gauge located directly downstream of a large (>30MW) reservoir and with $< 10\%$ of basin area impacted by large reservoirs as of 2013; and being located more than 50% (by area) within the eight-state study region boundary. The full set of 326 basins are those over which agricultural and forest cover summaries were calculated (see Figure 4.1).

On average in the region, agricultural land development began to plateau after 1990, and only a subset of basins in the region (for which we have data) experienced increases in agricultural land cover, synonymous with deforestation, through the 2000s (see Figure B.7). In general, agricultural land cover increases correspond to forest cover losses, except in cases of urbanization where agricultural land cover reductions correspond to no change or continued losses in forest cover. This was verified by observing agricultural and forest cover trends in basins over the same period of time (Figure B.3), and is the rationale for the use of only basins with less than 20% mean agricultural land cover prior to 1990 for any model fitting exercise (DID and mixed effects regression models).

We quantified deforestation and prior agricultural land use in river basins using 1km gridded annual agricultural land use (including cropland and pastureland, 1950-2012) [131] and 30m forest cover/loss (> 5m tree cover, 2000-2013) data [132]. We acquired forest cover and change data from Google Earth Engine (http://earthenginepartners.appspot.com/science-2013-global-forest/download.html) and generated annual forest cover from year 2000 forest cover and subsequent annual forest loss data (https://sites.google.com/site /earthengineapidocs/tutorials/global-forest-change-tutorial). The agricultural land cover dataset incorporates year 2000 forest cover from the forest change dataset, along with spatial reconstruction of historical agricultural census data [131]. Thus, the datasets are not independent, and it is expected that they provide similar results in river basins where agricultural land development is synonymous with forest removal.

We selected the monthly GPCC v.7 [154, 80] product to supplement in-situ rainfall data due to its long-term coverage (data prior to 1983) and least bias relative to in-situ rainfall data, as compared to other monthly datasets with similar temporal coverage (CRU TS3.10 and PREC/L .v1). We checked the consistency between monthly summaries of interpolated rainfall and GPCC (correlation = 0.95).

Forest change

An average of 45% of the cumulative area deforested in Amazon-Cerrado river basins between 2000-2013 was from high-density forests (areas with > 85% tree cover); 24% was from medium-density forest (areas with between 55-85% tree cover); and 32% was from what was likely savanna and/or pasture (areas with between 10-55% tree cover, after [182]). 19% of basins lost areas of high-density forest equivalent to between 5-34% of their total area, while the percent of medium-density forest and savanna/pasture land that was deforested in any basin was less than 5% of total basin area.

Flow normalization statistic

In our analyses, we used values of flow [m^3/s] normalized by basin-specific historical mean flow. This choice was based on an exploratory analysis of potential normalizing statistics that included the mean, median, and the 99th percentile. These statistics were chosen because they would produce normalized flow values that were interpretable at an individual basin scale, and the exploratory analysis determined the consistency of each statistic with respect to varying record durations. We took data from 176 river basins with more than 10 years of daily data and less than 10% missingness, and evaluated: (i) the temporal consistency and stationarity of each statistic by calculating variability and trends in one-year rolling-window estimates; and (ii) the sensitivity of statistics to record duration and bias by comparing the deviation between rolling-window and full-record estimates.

To look at variability and trends in the statistics over time (i), we calculated: (a) the coefficient of variation (rolling-window standard deviations as a fraction of rolling-window mean values); (b) simple linear slope estimates (linear regression of rolling-window statistic

values on time); and (c) resulting linear model predictions of total one-year change as a fraction of the rolling-window mean. On average across all basins evaluated, the coefficient of variation (a) was 30% for the median and 99th percentile, and 20% for the mean; positive and negative trends (b) were found in almost equal proportion (slightly more negative) for all statistics; and the (absolute) interquartile range (IQR) of total predicted annual change as a percent of the mean (c) was similar across all statistics, but slightly less for the mean (<1% for the mean versus >1% for the median and 99th percentile). On average across all basins, rolling-window deviations as a fraction of full-record mean values (ii) were - 0.5% for the mean, and +5% and -12% for the median and 99th percentile, respectively. Thus, the deviations and bias of the mean were least. In summary, the mean was determined to provide a normalization statistic that would respond most consistently across a sample of river basins for which record durations vary.

Diference-in-differences (DID) analysis

Prior to 1990, control basins tended to have low to mid-range (median) normalized flow rates that were on average higher than those of treatment basins; however, after 2000, treatment basin distributions of low (especially) and median flow were more similar to those of control basins, while high flow distributions remained relatively similar between groups in pre- and post-treatment periods (Figure B.5). This is the dynamic quantified by the DID regression.

DID linear fixed effects regression model

The DID model used in this study is:

$$log(flow_{imp} + 1) \sim \alpha + \gamma_i + \lambda_m + \beta_1 \text{rain}_{imp} + \beta_2 \text{temp}_{imp} + \delta 1(\text{post-treatment})_p + \tau 1(\text{treated})_i \cdot 1(\text{post-treatment})_p + \epsilon_{imp}$$

Observations in the above model are for basin i, in month of year m, and period p (pre- or post-treatment); $flow_{imp}$ is a flow percentile (i.e. 5th, 50th, 95th) calculated from daily flow observations in basin i in month m and period p, and normalized by basin mean historical flow percentiles; α is an intercept; γ_i is a basin fixed effect (estimated relative to the basin with the lowest average 50th percentile flow); λ_m is a month-of-year fixed effect (estimated relative to September, the lowest flow month in the study region); rain_{imp} and temp_{imp} are average rainfall (monthly total) and temperature anomaly (daily average) for basin i in month m and period p; $\delta 1(\text{post-treatment})_p$ is an indicator variable equal to one for observations in the post-treatment period (after 2000), and otherwise zero; $\tau 1(\text{treated})_i \cdot 1(\text{post-treatment})_p$ is the interaction of treatment basin and post-treatment period indicators, and estimates the treatment effect (on the treated); and ϵ_{imp} is an error term. The coefficient τ is the primary coefficient of interest, and measures the differential effect of the treatment (> 10% forest cover between 2000-2013) on the flow of a treatment group versus on a control group:

$$\hat{\tau} \;=\; (\overline{y}_{\text{treated, post}} - \overline{y}_{\text{treated, pre}}) - (\overline{y}_{\text{control, post}} - \overline{y}_{\text{control, pre}})$$

where $y = log(flow_{imp} + 1)$, and 'pre' and 'post' refer to the two time periods, before and after treatment. The effect of treatment (a difference of one unit in the period*treatment interaction term) can be approximated as $((e^{\tau} - 1) * 100)$.

For all flow percentile outcome specifications, this model meets ordinary least squares (OLS) regression assumptions based on a suite of common diagnostics, although model fit is poorer at low and high percentiles. The model fit according to the R^2 was between 0.75-0.81 for all percentiles. Regression diagnostics used to evaluate the models' underlying statistical assumptions included inspection of Normal quantile-quantile (QQ) plots, residual histograms, residual vs. fitted value plots, scale-location plots, and residual vs. leverage plots. We selected the log-linear model, wherein the outcome variable (flow) was offset by 1, based on model fit along the full range of flow percentiles, as well as for the log-linear model's ease of interpretation. Box-Cox transformations of the outcome variable would have provided better custom fits for each flow percentile regression (e.g. different transformations for the 5th vs. 95h percentile); however, the DID regression results were not significantly different, and such a transformation would have resulted in increased complexity in the interpretation of results.

Linear mixed effects regression model specification of the DID model

The complimentary mixed effects linear regression model is:

$$\begin{aligned} log(flow_{imp} + 1) \sim & [\alpha + (\alpha_j, \alpha_i)] + \lambda_m + \beta_1 \text{rain}_{imp} + \beta_2 \text{temp}_{imp} + \\ & \delta 1(\text{post-treatment})_p + \tau 1(\text{treated})_i \cdot 1(\text{post-treatment})_p + \epsilon_{imp} \end{aligned}$$

All terms are the same as those described for the traditional DID regression model. However, in this mixed effects model, α is a fixed intercept, and α_j, α_i are random intercepts, for nested groups and sites within nested groups.

DID assumptions

A key assumption of the DID model is that temporal trends in the treatment group's outcome (flow) would, absent treatment (deforestation), parallel trends in the control group. It is not possible to directly check this assumption; however, comparisons of pre-treatment trends in flow across treatment and control basins (Figure B.6) is suggestive of the validity of the assumption. Comparisons for different percentiles of flow, from low to high, show that linear temporal trends in normalized flow between the beginning of basin flow records (multiple) and 1990 are parallel for low to medium-range flow. However, trends for high rates of flow show differences between groups: the treated basin mean trend is increasing while the control

basin mean value is unchanging. The effect of this, according to DID model assumptions, is that the treatment effect is over-estimated at high flow. Given the high flow (>90th percentile) treatment effect was small and insignificant, this does not alter the interpretation of findings in this study

Matching

Matching is used to improve causal inferences in observational studies, and aims to reduce imbalance in the distributions of observed pre-treatment variables between treatment and control groups by subsetting data so that it is more (statistically) representative of data from a randomized experimemt [142, 159, 160]. Before selecting Mahalanobis Distance Matching (MDM) [158], we also explored Propensity Score Matching (PSM), another popular matching method, which produced similar DID results (selection of basins into final dataset had about 80% overlap between PSM and MDM), but we ultimately selected MDM due to evidence of superior performance of the MDM method in cases of limited data and a small number of covariates [183, 160].

MDM performance was evaluated in terms of its reduction in mean differences between covariates across treatment and control groups, and Student's t-tests of differences in means of those covariates across treatment and control groups. Matching reduced the mean differences between all treatment and control covariates by between 10-79%, and Student's t-tests failed to reject the null (of no difference in means) at even a 90% confidence level for all covariates except for pre-treatment % agricultural land cover (t-test p-value = 0.006). The imbalance of prior agricultural land cover indicates possible bias in the selection of basins into treatment and control groups with respect to agricultural land use (and previous deforestation) history. This is unavoidable given data limitations. Treatment basins are more likely to have experienced some level of agricultural development as early as the 1960s, which marks the beginning of the flow record. For this reason, only basins with less than 20% agricultural land cover prior to 1990 were selected into either the treatment or control groups, a threshold based on a previous finding that flow changes are generally insignificant when basin land cover change is under this threshold [128], and with the understanding that historical agricultural land cover indicates removal of natural vegetation [131, 1].

Because the DID method estimates the differential effect of treatment on a treatment group versus treatment on a control group, this imbalance may bias results if the effect (basin flow response) is different across levels of (< 20%) antecedent agricultural development. Figure B.7 shows that flow change (mean differences in percentiles of flow across periods) is not necessarily consistent across levels of pre-treatment agricultural land cover (% basin area) for treated basins. The greater the percentage of a basin's pre-treatment (historical) agricultural land cover, the lesser is the flow change response (less positive change, or smaller increases in flow) in treated basins from pre-treatment to post-treatment.

Treatment basins, as they are defined, tend to have greater pre-treatment agricultural land cover percentages; a reality imposed by data limitations (with respect to the duration of both flow and forest cover data). Prior to 1990, the majority of control basins have

agricultural land cover percentages (of basin area) that are less than 5%, while treatment basin agricultural land cover is spread more evenly between 2% and 20%. Based on graphical analyses (Figure B.7), flow change (increase) is greater for treatment basins that are less agriculturally developed prior to additional (2000-2013) deforestation. This suggests that evaluation of treatment effects on treatment basins with less antecedent agricultural development - a better matched analysis - would yield a larger treatment effect estimate.

Results

Table B.3 shows estimated DID model coefficients for the 5th, 50th, and 95th percentile flows; these are the same as those presented continuously in Figure 4.3 in the main text. Coefficients on the temperature terms indicate that the temperature anomaly, which captures deviation in average temperature from to a 1960-1990 baseline, has a weak negative relationship with flow. Rainfall has a significant, positive relationship with flow, as expected. The interpretation of the negative coefficient on the post-treatment indicator is that absent deforestation and holding all else constant, flow is actually decreasing with time on average. The regression coefficient estimates for rainfall and temperature include the combined effects of any exogenous climate forcing and climate-feedback effects from previous land use change, which may be significant [15, 16, 17, 20, 21].

Alternative specifications

A graphical summary of coefficient estimates with the specification from the main text (described in the Methods, and summarized in Table B.3), are provided in Figure B.8. Traditionally, an indicator for treatment would be included in the regression model. In the model described in the main text, the treatment indicator term is absorbed by the basin fixed effects (γ_i) term. We fit an alternative model that includes this term (instead of basins fixed effects). See Table B.4 for results. This alternative specification estimates similar treatment effects and other coefficients. The coefficient estimate on the treatment indicator shows that treated basins have lower flow on average than control basins at low and medium ranges of flow, which is also illustrated in Figure B.5. The linear mixed effects model includes basin and basin-in-nested-group random effects (to account for correlation in the observations of spatially overlapping basins), as well as a treatment indicator such as in the alternative specification described above and in Table B.4.

Streamflow volume change analysis

Linear mixed effects regression model

A chief benefit of mixed effect models, in addition to being able to account for correlation between basin observations, and basins within nested groups, is the capacity to represent basin-level trends in basins with more complete data, and to draw trend estimates in basins with sparse data towards the overall mean - what is referred to as ‘partial pooling’, which

can improve predictive capacity relative to traditional fixed effects models [152]. The mixed effects linear regression model used to estimate the relationship between mean-normalized annual total (depth) of streamflow and either forest or agricultural land cover is:

$$flow_{ijt} \sim [\alpha + (\alpha_j, \alpha_i)] + [\tau + (\tau_j, \tau_i)] \ \%\ \text{cover}_{it} + \lambda_t + \beta_1 X_{i1} + \beta_2 X_{it2} + \epsilon$$

Observations in the above model are for basin i in nested basin group j (basins with overlapping areas and/or located along the same river network) in year t; $flow_{ijt}$ is basin i (in nested group j) and year t flow depth (sum of daily volumetric flow normalized by basin area) as a percent of basin mean annual flow; α is a fixed intercept, and α_j, α_i are random intercepts, for nested groups and sites within nested groups; τ is a fixed slope, and τ_j, τ_i are random slopes, for nested groups and sites within nested groups with respect to % cover_{it}, which is the percent of basin i area forest cover that was lost, or agricultural land cover that was gained, in year t; λ_t is an annual time trend; X_{i1} are time-invariant basin physical features including top- and sub-soil (average) sand and clay percent and organic carbon (% weight), maximum basin elevation difference, and basin area; X_{it2} are basin climate covariates including rainfall (total) and temperature (daily average) for basin i and year t; and ϵ is an error term. t is the calendar year for land cover variables (January-December), and water year for flow and climate variables (October-September, starting in the same year as the calendar year).

The interpretation of the pooled, fixed-effect slope coefficient τ is that a 1 unit (percentage point) change in basin forest loss or agricultural land cover gain in a given year corresponds to a τ unit (percentage point) change in mean-normalized basin flow, holding all else constant; $\tau + \tau_j$ and $\tau + \tau_i$ provide estimates for individual basin nested groups and sites, respectively. Fitted or predicted values of percent of mean flow $\hat{flow}_{ijt}$ are converted to flow rates (mm/year) using the mean flow rate for each basin ($\hat{\text{flow}}_{ijt} *$ long-term mean $\text{flow}_i/100$).

Basin mean annual flow values for for the full set of 326 basins is based on the sum of average monthly totals constructed from mean daily flow by month across all available years of data for each basin, which allows calculation of annual totals even with significant missing flow values in some years. The RMSE between these estimated mean annual flow values and the actual mean annual totals from 226 basins with minimal missing data (fewer than 3 missing days per year, and ten or more years of data) is 17.5mm, which is 3% of basin mean annual totals on average. Predictions of annual flow (and change) on only basins used in model fitting (within-sample) generates roughly equivalent results as prediction across all basins.

Average annual flow change from land cover change for basin i in year t is $(\tau + \tau_i) *$ % cover change_{it}, where cover change is the loss of forest or gain of agricultural land cover in year t (relative to year $t-1$) for each year over the period of record (2000-2012 for forest, 1950-2012 for agricultural land cover). If the basin was not included in model fitting, $\tau_i = 0$, and the estimate is based on the fixed effect only. Model coefficient estimates were made by optimizing the restricted maximum likelihood (REML) criterion, p-values and degrees of

freedom are based on the Satterthwaite approximation, and full model (total flow change) and land cover (only) change predictions were estimated with 95% bootstrap prediction intervals and confidence intervals, respectively (see SI Text "Software" section).

Software

Analyses were carried out in the Comprehensive R Archive Network (CRAN) [73] programming environment (Version 3) on both Apple and Windows operating systems. Formatting and analysis of spatial data relied on the R spatial analysis packages `sp` [171, 172], `raster` [173], `gstat` [174], `geosphere` [184] and their dependencies; data manipulation relied on the `reshape2` [185] and `dplyr` [186] packages. Fixed effects linear regressions (DID analysis) relied on the standard R `stats` package, and linear mixed effects modeling relied on the `lme4` [187], `lmerTest` [188], and `merTools` [189]. Figures and maps were generated using using the `ggplot2` [179], `ggmap` [180], and `globe` [181] packages and their dependencies.

B.2 Figures

All figures were generated using the Comprehensive R Archive Network (CRAN) [73] programming environment (Version 3) on both Apple and Windows operating systems. See the SI Text "Data" section, and "Methods" section of the main text for data source references.

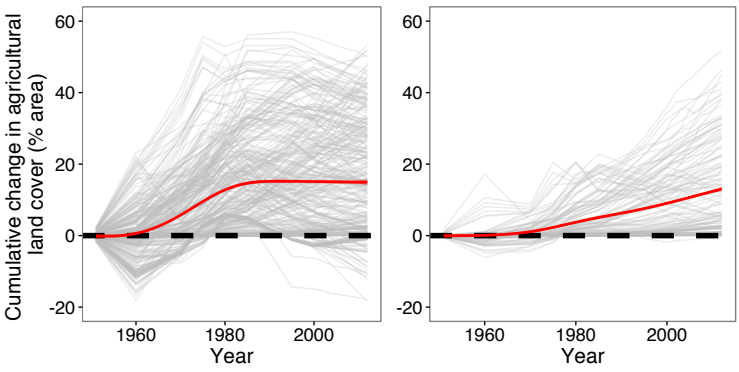

Figure B.1: Change in cumulative agricultural land cover over time

Change is annual cumulative change in the percent of basin area categorized as agricultural land cover (cropland and pastureland) between 1950-2012 [131]. Thin gray lines are individual basin time trends, and the thick red line is a local polynomial regression fit. The figure on the left shows all 326 basins in the study region, and the figure on the right shows 130 basins with $< 20\%$ agricultural land cover prior to 1990, which was a component of the criteria used to subset basins for the FDC and DID analysis.

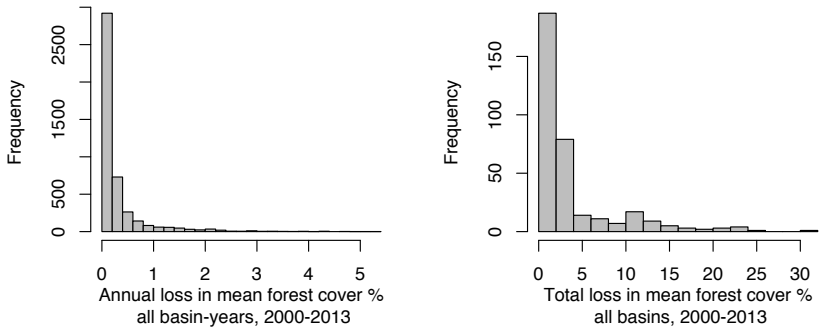

Figure B.2: River basin losses in forest cover

The mean forest cover percent is the basin-area mean of all (30m) pixel-level values of percent (>5m) tree cover [132]; losses are negative changes in the basin-area mean forest cover percent. A histogram of annual losses is shown on the left, including observations from all basins in all years between 2000-2013 (n=4238), and a histogram of total cumulative losses between 2000-2013 is shown on the right, with one observation per basin (n=326).

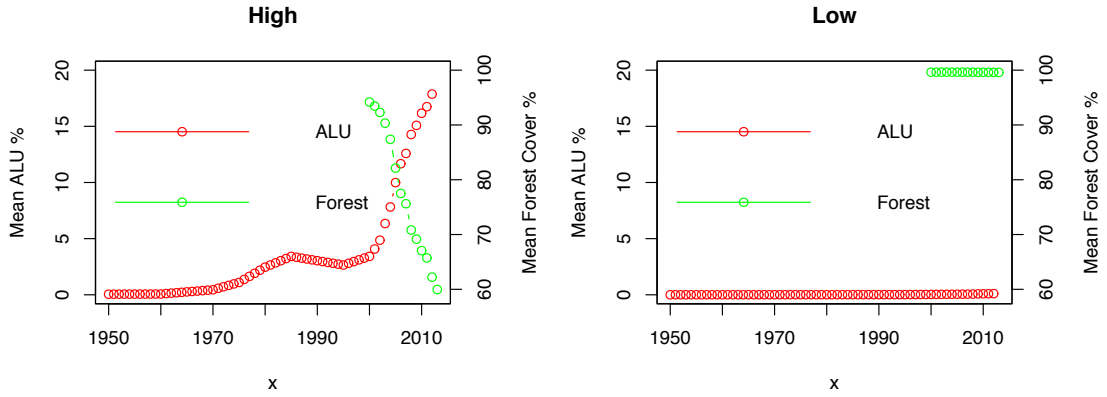

Figure B.3: Forest cover losses and corresponding agricultural land cover gain

The figures show annual agricultural land use (ALU) [131] (left axis, red) and forest cover [132] (right axis, green) as a percent of basin area in a high-deforestation river basin (left panel), and low-deforestation basin (right panel). Land cover change trends in these basins illustrate the correspondence that is present between forest and agricultural land cover in basins with $< 20\%$ agricultural land cover prior to 1990.

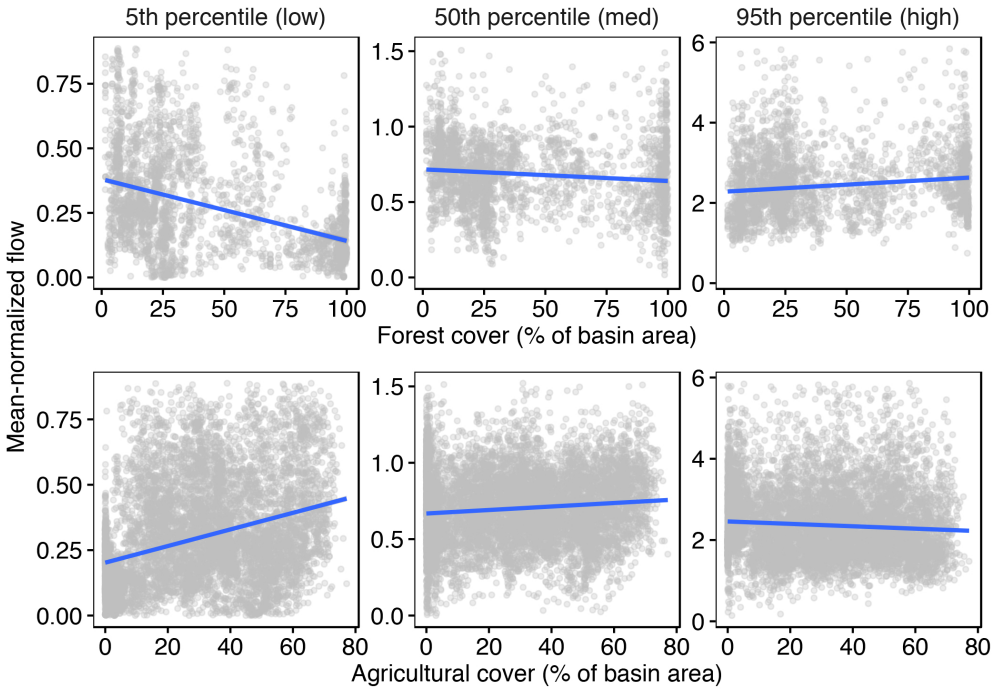

Figure B.4: Relationships between flow and land cover

Mean-normalized flow percentiles (5th, 50th, and 95th) are plotted against the percentage of total basin area categorized as forest (top) and as agricultural, including cropland and pasture (bottom). The blue trend line and 95% confidence interval are from robust linear regression. Each point represents a basin-year observation from basins with less than 10% missing values and fewer than 30 consecutive missing values (days). The forest cover data [132] is annual over the period of 2000-2013 (top: 2,579 basin-year observations of 323 basins with between 1-14 observations each); the agricultural land cover data [131] is annual over the period 1964-2012 (bottom: 7,533 basin-year observations of 326 basins with between 1-59 observations each).

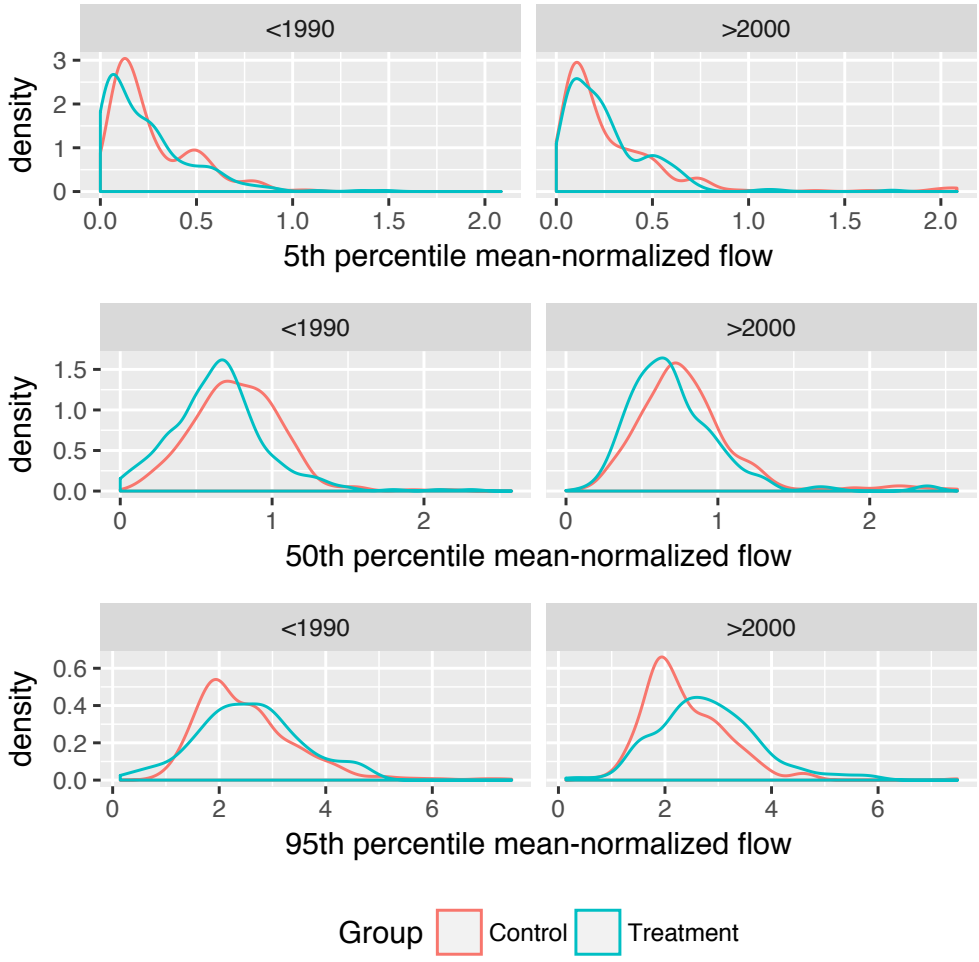

Figure B.5: Histograms of mean-normalized flow percentiles across periods and groups

Histograms are of 5th percentile (top), 50th percentile (middle), and 95th percentile (bottom) mean-normalized flow across treatment (high-deforestation) and control (low-deforestation) groups, which are indicated by color, and pre-treatment (<1990) and post-treatment (2000-2013) time periods, which are indicated by panel columns (left and right, respectively).

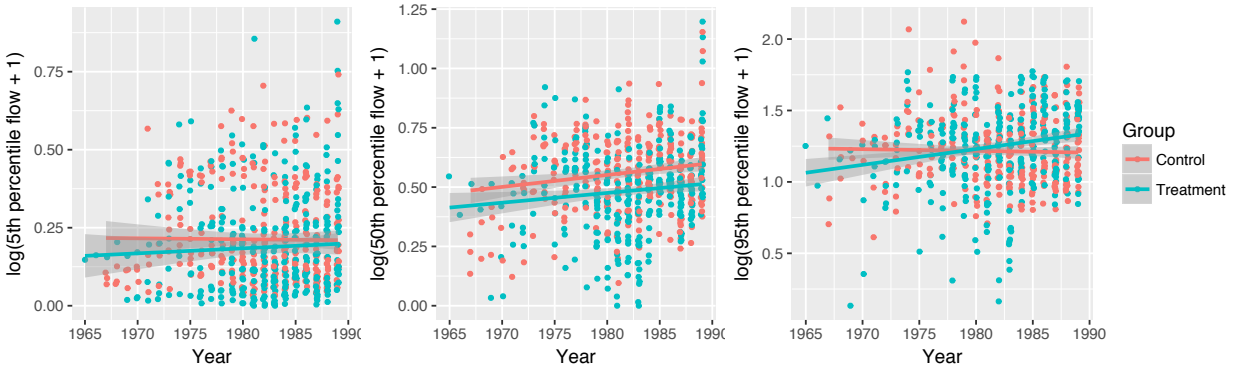

Figure B.6: Trends in pre-treatment (< 1990) flow by treatment and control group

The figures show scatterplots (points) of transformed (log and offset by one) normalized flow percentiles (vertical axis) along years prior to 1990; the lines and 95% confidence intervals are those from a robust linear regression; colors indicate high-deforestation and low-deforestation basin groups. The panels (left to rigth) show results for the 5th, 50th, and 95th percentile of normalized flow. Percentiles were calculated in all basin-years with less than 10% missing data for all basins (46) included in the DID analysis. Basin-years prior to 1990 are determined by data availability in each basin, and therefore vary.

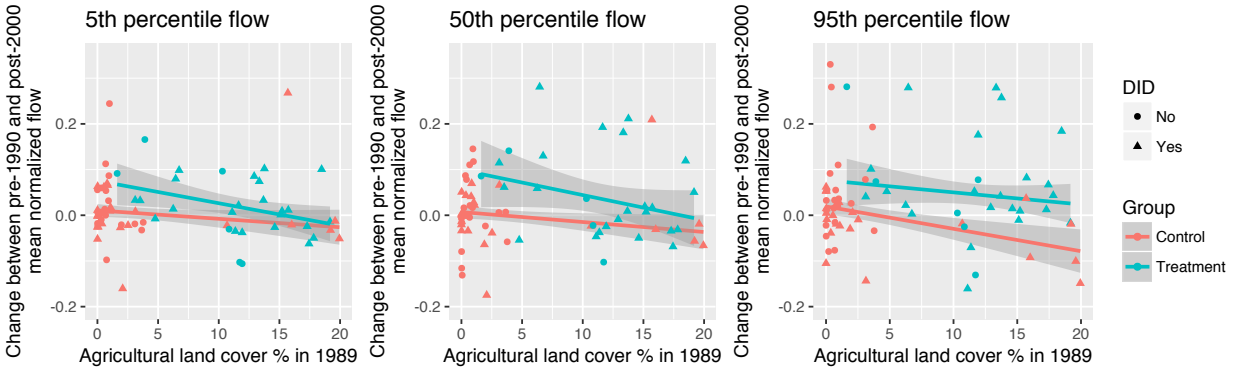

Figure B.7: Between-period flow change corresponding to pre-treatment agricultural land development

The vertical axes are change between pre-treatment (<1990) and post-treatment (>2000) period-mean (log-transformed) flow percentiles for the 5th (left), 50th (center), and 95th (right) percentiles of normalized flow. The horizontal axis is pre-treatment (1989) agricultural land cover as a percent of basin area. Points are values of change in individual basins with respect to each basin's level of pre-treatment agricultural development; shape indicates whether or not the basin was included in the DID analysis; colors indicate treatment group assignment of those basins; the lines and 95% confidence intervals are from a robust linear regression of change on agricultural land cover %. Results from 75 basins are shown, and include the 46 used in the DID analysis; additional basins that met data quality requirements for the DID analysis (other than matching criteria) were included here in order to best approximate the relationship between flow change and prior agricultural land use.

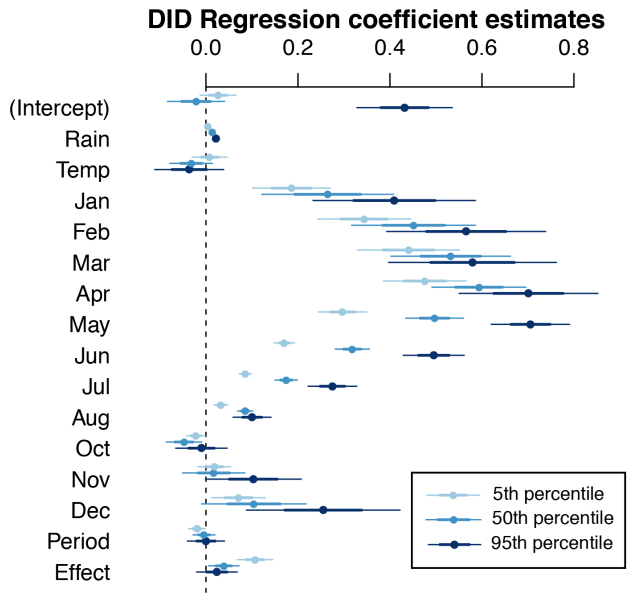

Figure B.8: Visualization of DID regression coefficient estimates

DID regression coefficient estimates for the 5th, 50th, and 95th percentile, including the intercept and month fixed effects (relative to a September baseline). All basin fixed effects (45, relative to the lowest flow basin), are not included in the graphic above, but all are significant at a 95% confidence level for all flow percentile regressions. The significance in monthly trends reflects seasonality in flow across the region.

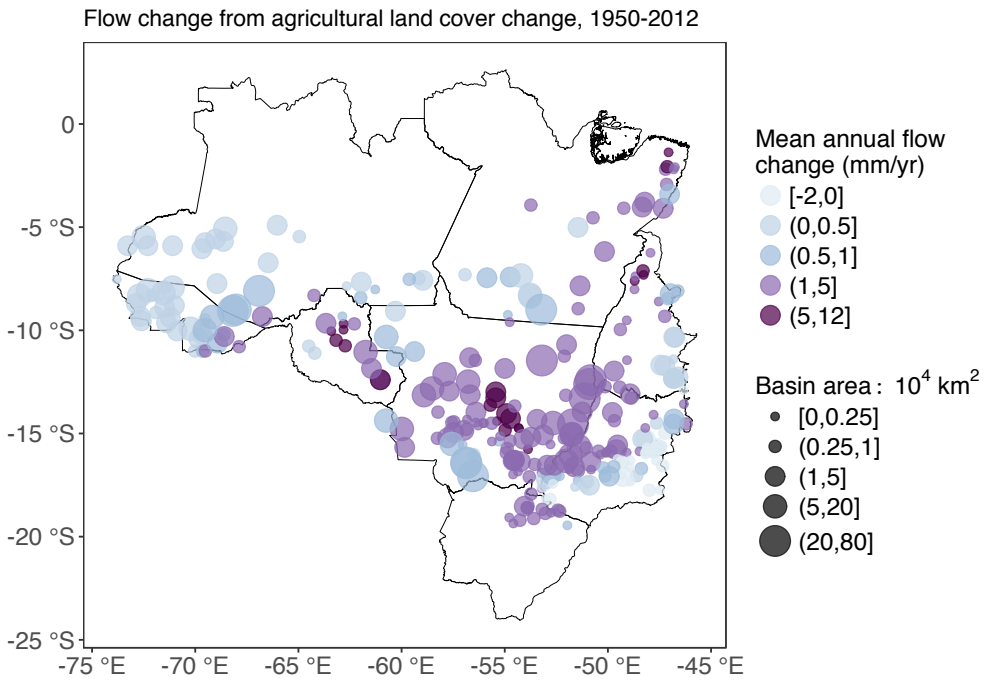

Figure B.9: The estimated effects of agricultural land development on annual flow

The map shows river basin centroids (n=326 river basins) [32]; colors indicate annual flow increases (mean mm/year, 1950-2012) due to agricultural land cover gain [131]; circle sizes indicate the size of the river basin. Values are basin mean estimates, and error bars are 95% confidence intervals. Estimates are mean predicted values of change calculated from mixed effects model coefficient estimates (see Methods).

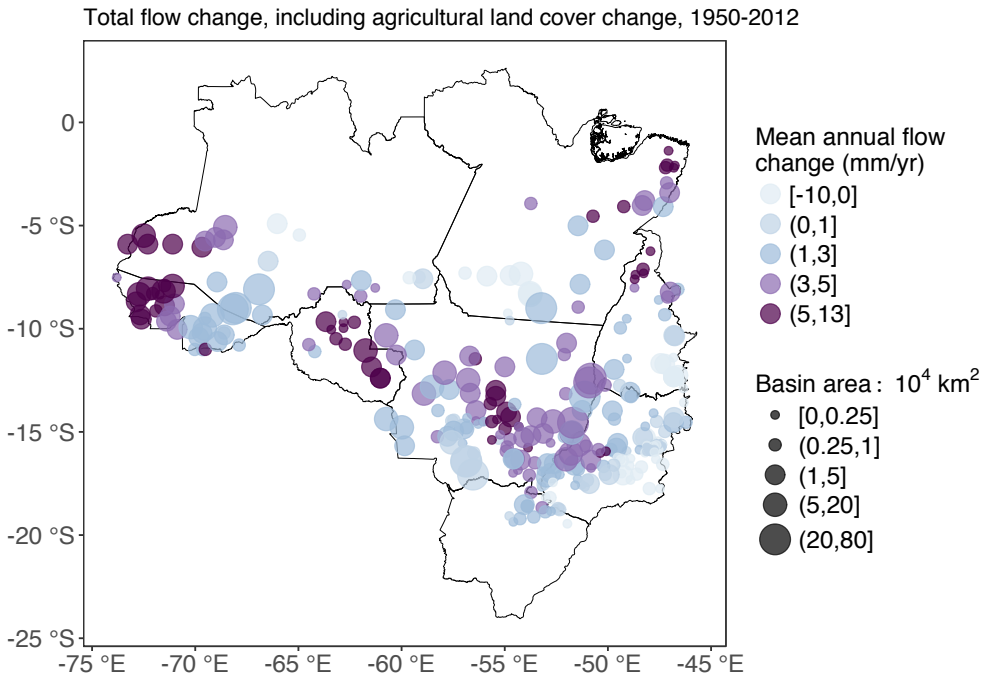

Figure B.10: The estimated effects of all environmental change, including agricultural land cover gain, on annual flow

The map shows river basin centroids (n=326 river basins) [32]; colors indicate annual flow increases (mean mm/year, 1950-2012) due to all environmental change (climate and agricultural land cover); circle sizes indicate the size of the river basin. Values are basin mean estimates, and error bars are 95% confidence intervals. Estimates are mean predicted values of change calculated from mixed effects model coefficient estimates (see Methods).

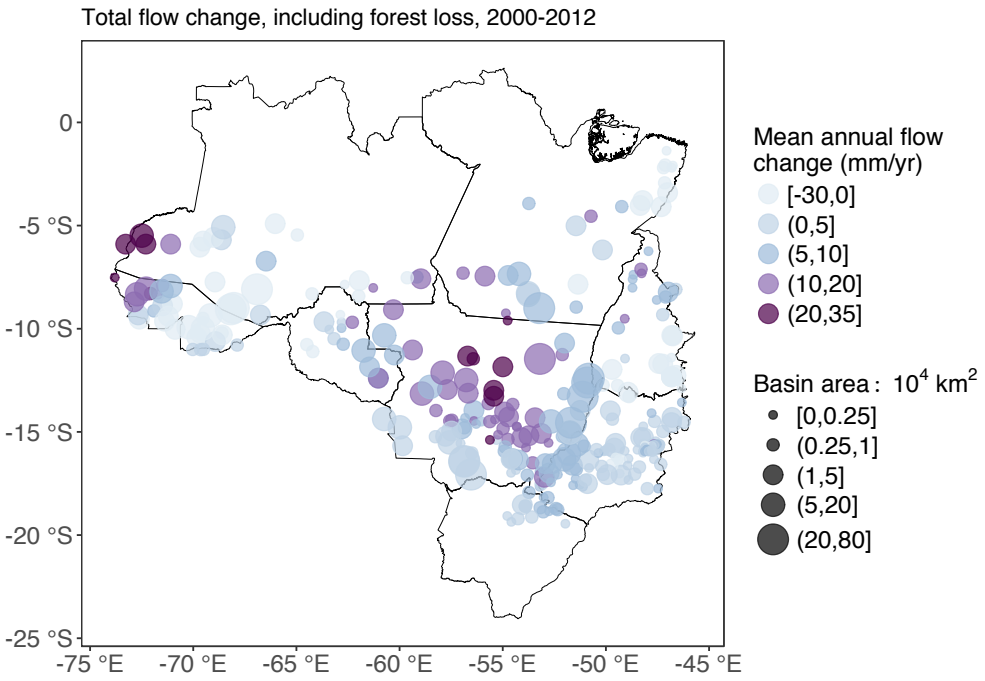

Figure B.11: The estimated effects of all environmental change, including forest loss, on annual flow

The map shows river basin centroids (n=326 river basins) [32]; colors indicate annual flow increases (mean mm/year, 2000-2012) due to all environmental change (climate and forest cover); circle sizes indicate the size of the river basin. Values are basin mean estimates, and error bars are 95% confidence intervals. Estimates are mean predicted values of change calculated from mixed effects model coefficient estimates (see Methods).

B.3 Tables

Table B.1: Data types, temporal resolution and duration, spatial resolution, and sources

Data Type	**Temporal Resolution**	**Duration**	**Spatial Resolution**	**Source**
River basin boundaries	static	-	polygon	Derived from Agência Nacional de Águas (ANA) data [32]
Streamflow (at basin outlets)	m^3/s	1980 - 2013	point	Derived from Agência Nacional de Águas (ANA) data [32]
Large reservoir (>30MW) locations and upstream drainage areas	static	2013	point, polygon	Derived from Agência Nacional de Águas (ANA) and Agência Nacional de Energia Elétrica (ANEEL)data [32]
Rainfall (in-situ)	mm/day	1983-2013	0.25 degree	Derived from Agência Nacional de Águas (ANA) data [32]
Rainfall (gridded)	mm/day	1983-2013	0.25 degree	NOAA Climate Data Record (CDR) of Precipitation Estimation from Remotely Sensed Information using Artificial Neural Networks (PERSIANN-CDR), v.1 [89]
Rainfall (gridded)	mm/month	1950-2013	0.5 degree	Global Precipitation Climatology Centre (GPCC) v.7 [154, 80]
Temperature (gridded)	C/day (mean or max anomaly and climatology)	1950-2013	1 degree	Berkeley Earth Surface Temperature (BEST) [156]
Soil	static	-	various (polygon)	Harmonized World Soil Database (HWSD) [157]
Elevation	static	-	90m	NASA Shuttle Radar Topography Mission (SRTM) v.4 [52]
Agricultural land cover	% agricultural cover/year	1950-2012	1km	Agricultural land use in Brasil (1940-2012) [131]
Forest cover and change	% forest cover/year	2000-2012	30m	Hansen/UMD/Google/USGS/NASA [132]

Table B.2: Data summary of Amazon-Cerrado river basin features

	Mean	Min	Max	SD
Mean annual flow (mm/year)	640.93	25.76	1867.01	279.29
First year in flow record	1982	1950	2009	14
Last year in flow record	2009	1968	2013	7
# years in flow record	27	2	64	14
Missingness of flow record (%)	0.1	0	0.85	0.12
Annual rainfall (mm)	1737.66	1295.89	2719.39	314.89
Mean daily temperature (C)	24.95	22.12	27.74	1.55
Soil sand (%)	46.06	3.5	91.5	18.33
Soil clay (%)	37.37	5	76.5	14.88
Soil organic carbon (% weight)	0.85	0.21	2.3	0.31
Mean elevation (m)	503.06	55.1	1228	296.91
Elevation change (m)	468.76	38	1277	237.84
Area (10^4 km^2)	2.68	0	44.75	6.27
Forest cover (mean % area, 2000-2012)	39.77	1.52	99.93	33.31
Forest cover change (mean % point/year, 2000-2012)	0.3	0	2.66	0.43
Agricultural land cover (mean % area, 1950-2012)	24.43	0	64.68	18.37
Agricultural cover change (mean % point/year, 1950-2012)	0.24	-0.3	0.86	0.26
# Basins	326	-	-	-
# Nested basin groups	43	-	-	-

Mean, minimum, maximum, and standard deviation of river basin features. Soil features are averages of topsoil and subsoil values, and are area-weighted averages over basin areas. Forest cover is the basin areal-average, annual percentage of basin area with tree cover > 5m in height according to 30m resolution forest cover and change data [132]. Agricultural cover is the basin areal-average, annual percentage of basin area with 1km resolution land classified as natural pasture, planted pasture, and annual and perennial cropland [131]. Climate data area basin areal-average, annual values. See Methods, SI Text "Data" section, and Table B.1 for complete data details and sources.

Table B.3: DID regression model estimates

	5th (Low)	50th (Median)	95th (High)
Effect	0.107 (0.019)***	0.039 (0.017)*	0.024 (0.022)
Rain	0.004 (0.002)**	0.014 (0.002)***	0.022 (0.004)***
Temp	0.008 (0.019)	−0.032 (0.023)	−0.036 (0.038)
Post	−0.019 (0.009)*	−0.004 (0.012)	−0.000 (0.020)
R^2	0.78	0.81	0.80

$^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$

The columns are separate regression models with the outcome variables of 5th, 50th, and 95th percentile mean-normalized flow. "Effect" is the estimated treatment effect of high-levels of deforestation on flow. "Rain" and "Temp" are the coefficients on average monthly-total rainfall (cm) and average daily-mean temperature anomaly in month-periods; "Post" is the post-treatment period indicator. An intercept, site fixed effects, and month fixed effects, were included in each regression. Standard errors (in brackets) are clustered by site. Each regression has 60 (parameters, including those in the table, the intercept, and 45 basin and 11 month fixed effects) and 1,043 (residual) degrees of freedom.

Table B.4: DID regression model estimates, alternative specification

	5th (Low)	50th (Median)	95th (High)
Effect	0.107 (0.019)***	0.038 (0.020)	0.023 (0.029)
Rain	0.002 (0.001)	0.012 (0.001)***	0.015 (0.001)***
Temp	0.002 (0.026)	−0.022 (0.027)	−0.038 (0.039)
Treated	−0.075 (0.014)***	−0.016 (0.014)	−0.017 (0.020)
Post	−0.018 (0.017)	−0.009 (0.017)	−0.002 (0.025)
R^2	0.61	0.79	0.74

$^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$

The columns are separate regression models with the outcome variables of 5th, 50th, and 95th percentile mean-normalized flow. "Effect" is the estimated treatment effect of high-levels of deforestation on flow. "Rain" and "Temp" are the coefficients on average monthly-total rainfall and average daily-mean temperature anomaly in month-periods; "Treated" is the treatment group indicator; "Post" is the post-treatment period indicator. An intercept and month fixed effects were included in each regression. Each regression has 16 (parameters, including the intercept) and 1,087 (residual) degrees of freedom.

Table B.5: Fixed effects estimates from fitted mixed effects model

	Forest	Agriculture
Forest	-0.64 $(0.86)^{***}$	
Agriculture		1.07 $(0.32)^{***}$
Rain	74.69 $(2.62)^{***}$	56.75 $(1.54)^{***}$
Temp	0.27 (0.91)	-6.54 $(1.15)^{***}$
Sand	-0.04 (0.09)	-0.00 (0.14)
Clay	-0.11 (0.13)	-0.20 (0.17)
OC	12.58 $(4.70)^{**}$	8.11 (4.72)
Elevation	1.60 $(0.44)^{***}$	1.75 $(0.43)^{***}$
Area	-0.17 (0.12)	-0.19 (0.15)
Year	0.12 (0.12)	0.04 (0.05)
Num. obs.	1410	2016
Num. basins	173	91
Num. nested basin groups	22	22

$^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$

The columns include fixed effects coefficient estimates from separate linear mixed effects regression models with predictor variables of (left) forest cover between 2000-2012 and (right) agricultural land cover between 1950-2012 (% basin area/year). "Rain" and "Temp" are the coefficients on average annual-total rainfall (m) and average annual daily-mean temperature (degrees C); "Sand" and "Clay" are the coefficients on sand and clay percents in topsoil and subsoil combined; "OC" is the coefficient the organic carbon content in topsoil and subsoil combined (% by weight); Elevation is the coefficient on the maximum basin elevation change (m); "Area" is the coefficient on the area of each basin (10^4 km^2), and "Year" is the year of observation. All variables were area-weighted averages over basin areas. Standard errors are in brackets. Random intercepts and slopes with respect to land cover were estimated for each nested group and basin (see Methods).