Title
Bridging the Information Gap: Remote Sensing and Micro Hydropower Feasibility in Data-Scarce Regions

Author
Muller, Marc Francois

Publication Date
2015

Peer reviewed|Thesis/dissertation

Bridging the Information Gap: Remote Sensing and Micro Hydropower Feasibility in Data-Scarce Regions

by

Marc François Müller

A dissertation submitted in partial satisfaction of the

requirements for the degree of

Doctor of Philosophy

in

Engineering – Civil and Environmental Engineering

in the

Graduate Division

of the

University of California, Berkeley

Committee in charge:

Professor Sally Thompson, Chair
Professor Slav Hermanowicz
Professor Maggi Kelly

Spring 2015

Bridging the Information Gap: Remote Sensing and Micro Hydropower Feasibility in Data-Scarce Regions

Abstract

Bridging the Information Gap: Remote Sensing and Micro Hydropower Feasibility in Data-Scarce Regions

by

Marc François Müller

Doctor of Philosophy in Engineering – Civil and Environmental Engineering

University of California, Berkeley

Professor Sally Thompson, Chair

Access to electricity remains an impediment to development in many parts of the world, particularly in rural areas with low population densities and prohibitive grid extension costs. In that context, community-scale run-of-river hydropower – *micro-hydropower* – is an attractive local power generation option, particularly in mountainous regions, where appropriate slope and runoff conditions occur. Despite their promise, micro hydropower programs have generally failed to have a significant impact on rural electrification in developing nations. In Nepal, despite very favorable conditions and approximately 50 years of experience, the technology supplies only 4% of the 10 million households that do not have access to the central electricity grid. These poor results point towards a major information gap between technical experts, who may lack the incentives or local knowledge needed to design appropriate systems for rural villages, and local users, who have excellent knowledge of the community but lack technical expertise to design and manage infrastructure. Both groups suffer from a limited basis for evidence-based decision making due to sparse environmental data available to support the technical components of infrastructure design.

This dissertation draws on recent advances in remote sensing data, stochastic modeling techniques and open source platforms to bridge that information gap. Streamflow is a key environmental driver of hydropower production that is particularly challenging to model due to its stochastic nature and the complexity of the underlying natural processes. The first part of the dissertation addresses the general challenge of Predicting streamflow in Ungauged Basins (PUB). It first develops an algorithm to optimize the use of rain gauge observations to improve the accuracy of remote sensing precipitation measures. It then derives and validates a process-based model to estimate streamflow distribution in seasonally dry climates using the stochastic nature of rainfall, and proposes a novel geostatistical method to regionalize its parameters across the stream network. Although motivated by the needs of micro hydropower design in Nepal, these techniques represent contributions to the broader international challenge of PUB and can be applied worldwide. The economic drivers of rural electrification are then considered by presenting an econometric technique to estimate the

cost function and demand curve of micro hydropower in Nepal. The empirical strategy uses topography-based instrumental variables to identify price elasticities.

All developed methods are assembled in a computer tool, along with a search algorithm that uses a digital elevation model to optimize the placement of micro hydropower infrastructure. The tool – Micro Hydro [em]Power – is an open source application that can be accessed and operated on a web-browser (http://mfmul.shinyapps.io/mhpower). Its purpose is to assist local communities in the design and evaluation of micro hydropower alternatives in their locality, while using cost and demand information provided by local users to generate accurate feasibility maps at the national level, thus bridging the information gap.

A papa...

Contents

List of Figures

List of Tables

Acknowledgments

This dissertation would not have been possible without the loving support of my "tribe", Michèle and Micah Müller. My wife Michèle is the smartest person I know – her math wizardry contributed to every single chapter of this dissertation in one way or another, and has prevented an untold number of embarrassing mistakes. Thanks for coping with my crazy self-imposed deadlines and thanks for the countless hours spent debating my research around the kitchen table. You have been the ideal partner through my PhD: a perfect *vis-à-vis*. Thanks to my son Micah, whose constant joyfulness and wonder are my greatest inspiration. Spending so much time with you was essential to my sanity and is the most cherished memory of my time at Berkeley. Je vous aime de tout mon coeur!

I also wish to thank my mother, Yuthor Müller, for her support during all these years. Thanks for your constant encouragement and for instilling in me the curiosity and dedication that brought me into science and engineering. Behold: I finally did become a doctor!

The research contained in this thesis was conducted under the stellar supervision of Prof. Sally Thompson. Sally is without a doubt the best research advisor a graduate student could wish for! Thank you so much for your time, dedication and responsiveness – your input was instrumental in framing my ideas into publishable research. I am also grateful for my committee members and advisors at Berkeley: Slav Hermanowicz, Maggi Kelly, Ashok Gadgil, Kara Nelson and Isha Ray – thanks for your unwavering support!

Field work in Nepal was made possible through the support of HELVETAS Swiss Intercooperation. Thanks to Agnes Montangero, Madan Bhatta, Mohan Batta, Bikram Rana and Moritz Güttinger for the amazing logistical support on the field. The research also benefitted from the expert advice of Jürg Merz, Sumanta Neupane, Sridhar Devkota and Giriraj Adhikari in Nepal, and Martin Bölli, Alex Arter, Fred Jordan and Philippe Heller in Switzerland and Ted Miguel, Fenella Carpena and Tadeja Gracner at Berkeley. Thanks to David Dralle, Morgan Levy and Gabrielle Boisramé from the Thompson lab at Berkeley for their edits and valuable feedback on individual chapters of this dissertation.

My research was generously supported by the US Department of State through the Fulbright Science and Technology Award, by the Swiss National Science Foundation through the Doc.Mobility fellowship program, and by Berkeley's Big Ideas competition.

Finally, my greatest thanksgiving goes to God – the reason I'm here in the first place. Thanks for Your guidance and limitless support!

Chapter 1

Introduction

Water and energy are essential to both human well-being and to economic development. They are traditionally related in a paradoxical nexus: Energy production consumes significant amounts of water; providing water, in turn, consumes energy [91]. Unfortunately, both remain inaccessible to millions of people, particularly in rural areas of developing countries [112, 236], where financial barriers often impede the development of rural water supply schemes and the extension of centralized power grids [42].

To avoid grid extension costs, distributed power generation through micro run-of-river hydropower infrastructure can be commissioned to supply isolated and locally owned micro-grids in mountainous areas. This approach has been implemented in Nepal for the past 50 years to complement the centralized power grid and has a documented ability to generate net profit in rural developing regions if appropriately implemented [128, 122]. Micro-hydropower also offers interesting opportunities to reverse the water-energy nexus and foster access to drinking water, as seen in Switzerland, where micro-hydropower plants have been integrated into gravity-fed water supply networks for the last hundred years. About 80 GWh/yr of electrical energy is generated from water supply infrastructure using this technology. Net profit from electricity is used to cross-subsidize water supply [111], improving the financial sustainability of water utilities. In developing countries, such cross-subsidies could support the development of driking water supply [227].

The feasibility of incorporating electricity generation into water supply systems (or vice-versa) remains largely unknown. Successful multi-use systems consisting of domestic water supply and income-generating irrigation water supply have been implemented within a single infrastructure [181, 148]. There is a need to explore the value of linking water and electricity supply within a single run-of-river operation in rural, developing regions. Unfortunately, pilot projects, which would inform on the suitability of the infrastructure in specific communities, have a debatable external validity and the capital intensiveness of the infrastructure causes high opportunity costs if the pilot experiment fails. This risk is particularly significant in remote areas of developing countries, where environmental and economic data are missing to inform economic and engineering design decisions.

Sustainable infrastructure design and installation in developing nations is challenged by

institutional barriers and decision-making norms. Policy and design decisions are typically made by technical experts, who may lack incentives or local knowledge needed to make a design appropriate to the local context. This lack of sensitivity can have disastrous long-term equity and welfare consequences [78]. In contrast, community members have excellent knowledge of the socio-economic and institutional context of the community, but typically lack the quantitative data and technical expertise to site, design and manage infrastructure. Both groups can suffer from a limited basis for evidence-based decision making due to the limited availability of environmental data needed to support the technical components of infrastructure design. The goal of this dissertation is to draw on recent advances in remote sensing data, stochastic modeling techniques and open source platforms to address that gap.

The research focuses on Nepal, which serves as a case study. There, despite favorable environmental and institutional conditions, domestic manufacturing capability and a mature economic sector, micro-hydropower has had a limited impact on rural electrification. Paradoxically, the extreme topography of the region is at once responsible for its enormous hydropower potential, and the low physical accessibility of most communities. I argue in Section 1.1 that the limited impact of micro-hydropower in Nepal is closely related to remoteness and poor connectivity, which have hindered the dissemination of the necessary knowledge for local communities to exploit the abundant hydropower potential often available in their backyard. Although the increasing availability of information and communication technologies has the potential to partly address this issue, the scarcity of quantitative local field data, from which to extract useful design information remains a key obstacle, as argued in Section 1.2. In fact, making Predictions in Ungauged Basins (PUB) remains a major global contemporary challenge in water resources[32]. Whilst much recent research has been devoted to predicting policy-relevant information using remote sensing data and advanced statistical models [see 32, for an extensive review], few (if any) are applicable to the extreme topography, hydrology and data-scarcity that are characteristic of Nepal, and of numerous other developing countries for that matter. I address these gaps in this dissertation by developing and validating globally applicable approaches to predict key information on the hydrology and economy of ungauged communities. To enhance their practical impact, these methods are integrated in a tool intended to automatize the placement and design of village scale micro-hydropower schemes and evaluate their economic feasibility. The tool can be accessed on the web and its use requires very little technical knowledge. It is available for local communities to assess the costs and feasibility of micro-hydropower and compile a large scale feasibility map to inform rural policy making. This bidirectional flow of information will hopefully contribute to bridging the information gap.

1.1 Mico-hydropower in Nepal

With a per capita gross national income of 750USD [1], Nepal is among the poorest countries in the world. It ranks 157th out of 187 countries in Human Development Index [1] and is in the process of recovering of a 10-year violent civil conflict that ended in 2006 [1]. Nonetheless,

thanks to its extreme topography and favorable hydrology, the country has one of the largest hydropower resources in the world with an estimate 83 GW of hydropower potential [1]. Yet most of the available resource remains untapped and the run-of-river[1] hydropower plants feeding the national grid are insufficient to meet peak demand during the dry season, leading to severe load shedding of up to 16 hours per day [1] affecting (mostly urban) grid connected households. More than 80% of the population lives in rural areas, 66% of which do not have access to grid electricity at all [141] due to high grid extension costs brought about by the challenging topography. As a result, less then half the country's population had access to electricity in 2009. Poor reliability and access to power are seen as the most serious infrastructure bottleneck to growth [1].

In that context, off-grid micro-hydropower schemes, defined in Nepal as run-of-river hydropower generating less than 100kW of electricity, stand out as cost-effective solutions for rural electrification. Their low environmental impact, their ability to stimulate the local economy and their effect on living standards and gender parity through income generating activities has been documented in numerous successful cases [e.g., 247, 87, 165, 241]. Nepal has a long tradition of extracting energy from flowing water, with traditional wooden turbines – *ghattas* – being used for centuries to process grain. Yet impetus to develop modern micro-hydropower really began in the 1970's, when high fuel prices and rising demands for rice hulling and flour milling lead to the development of metallic water wheels that increased the available shaft power by an order of magnitude. Such improved *ghattas* were able to compete with diesel mills on price and attracted a significant flow of capital, both from the private sector and international development agencies, which jump-started a domestic manufacturing industry for water turbines. Small electric generators were then gradually added onto existing milling systems through the mid 1980's, when a series of technical innovations and favorable institutional and financial policies boosted the dissemination of stand-alone micro-hydropower systems for rural electrification. Since then, an estimated 2900 micro-hydropower schemes have been installed under a large variety of institutional arrangements ranging from community-owned and operated public infrastructure to (mostly local) private ventures. Micro hydropower currently supplies about 350,000 rural households with about of 28 MW of electricity [164].

The relative success of micro-hydropower development in Nepal can be associated to a combination of key institutional and technical innovations. From a technical standpoint, the development of low-cost and robust electronic load controllers to replace traditional mechanical controls substantially decreased the cost of power generation [85]. Accessibility constraints also required the adaptation of turbine designs to local conditions: turbine components had to be manufacturable locally[2] and transportable on site by human porters on steep mountain trails[3]. This resulted in the establishment of a strong industry base capable

[1] No conventional hydropower dams with significant storage capacity is available in Nepal at the time of this writing.

[2] For instance, Nepal does not have any advanced foundry capability: Crossflow turbines are manufactured from steel sheet and pelton turbines are made of (less durable) bronze instead of steel.

[3] Such transportability requirements impose severe restriction on the weight of each component.

of developing, manufacturing, assembling and maintaining all micro-hydropower components (except generators) domestically[151]. Key institutional developments include the delicensing of all schemes below 100kW, the creation of a dedicated government micro-hydro credit program with partial subsidy of electrification equipment, and the creation of the Alternative Energy Promotion Center (AEPC), an umbrella governmental organization to promote, evaluate, subsidize and support micro-hydropower for rural electrification.

Yet despite a sustained growth for the last 50 years, micro-hydropower suffers from a high failure rate (up to 30% of existing plants are not in operation according to Khennas et al. [128]) and difficulties to scale [122, 151]. The 350,000 supplied households still represent a small fraction of the estimated rural population of about 10 million with no electricity access. Key challenges put forward in the literature include:

1. **Productive end use.** The ability to use the generated power to produce a marketable good is a strong predictor of the infrastructure's financial sustainability [128, 151, 87]. Yet a key constant preventing local enterprise development is the absence of information on financing sources and market prices [87]

2. **Local capacity building.** Technical and managerial knowledge is concentrated in urban manufacturing centers and challenging to access for remote communities [85, 87]

3. **Participative planning.** Local communities have little involvement in site selection and early design decisions. Centralized institutions typically make inventories of potential sites before "*attempting to persuade the local community that a micro-hydro electrification project is what it needs.*" [151]

4. **Input data.** Despite the hydrologic complexity of the region, streamflow measurement and regionalization techniques are very rudimentary[4] and notoriously unreliable. Streamflow is often overestimated, which, combined to overly optimistic demand forecasts, are frequent causes for over-designed and financially unsustainable infrastructure [151].

5. **Conflicts of interest.** It is normal practice in Nepal for turbine manufacturers to perform site surveys and feasibility studies [151], and for financing institutions to provide technical assistance to micro hydropower operators [128]. Both practices are liable to cause conflicts of interests and lead to poor designs (manufacturers have the incentive to oversize turbines) and biased financial assessments.

Although not exhaustive, the above list highlights the importance of local participation at all stages of micro-hydropower infrastructure development. Unfortunately in Nepal, de-

[4]The micro-hydropower association of Nepal recommends using the medium irrigation project method [60], which divides the counry into seven hydrological regions with predefined sets of monthly flow coefficients. Streamflow distribution is obtained by indexing the appropriate monthly coefficients with *one* dry season flow observation in the considered stream.

spite relatively inclusive and decentralized institutions[5], local participation is hindered by poor transportation infrastructure[6]. The ensuing physical remoteness obstructs the proper dissemination of technical knowledge and prevents the necessary local capacity from being built.

1.2 Dissertation Outline: Addressing The Information Gap

In an international assessment of best practices for micro-hydropower development, Khennas et al. [128] emphasized the need for a *technical catalyst*: a project developer with the technical expertise, tenacity and deep knowledge of the institutional and social fabric of the community to champion the development the project, while being able to work in close proximity to villagers at relatively low cost. A telling example of the impact of such local enablers can have is recounted in Ghale et al. [87]. Without a doubt, there is a large number of entrepreneurial community members in rural Nepal that have the tenacity and community knowledge to champion the development of successful micro-hydropower infrastructure, when provided with proper technical assistance. Yet, even equipped with web-connected smartphones, would be local catalysts face a series of prohibitive barriers related to the absence of quantitative local field data. The remainder to this section outlines some of these barriers and the approaches used in this dissertation to address them.

1.2.1 Accessing relevant data

Hydropower potential is driven by two environmental constraints, elevation difference and streamflow distribution, both of which can be directly or indirectly extracted from satellite data at no costs. For instance, the ASTER GDEM digital elevation model provides a global raster of land surface altitude at a 30 meter resolution, and the tropical rainfall measurement mission (TRMM) 3B42 product provides nearly two decades of near-global daily precipitation at 0.25 °($\approx$ 30km) resolution. Retrieving these datasets and extracting locally relevant information requires significant technical knowledge. In particular, rainfall estimates, which drive streamflow distribution, are notoriously biased in mountains due to the sensor's inability to properly detect orographic precipitation. Current adjustment algorithms do not correct errors in rainfall frequency and fail to properly account for spatial heterogeneities in the biases, both of which have a significant effect on streamflow generation in Nepal. I address these shortcomings in Chapter 2 by developing a novel bias adjustment procedure

[5]Locally elected *Village Development Committees* are responsible for service delivery and ensures proper use and distribution of state funds at the local level.

[6]Poor physical connectivity is seen as Nepal's second most important infrastructure challenge, with over one-third of the people in the hills living more than four hours away from an all-weather road, and 15 out of 75 district headquarters yet to be connected by a road [1].

that independently adjusts the parameters of a stochastic rainfall model and uses geostatistical regionalization and scaling techniques to resolve spatial heterogeneity. The algorithm was successfully validated in Nepal, where it outperformed NASA's ground-based calibration procedure. Chapter 2 was published as a research article (co-authored with Sally Thompson) in Advances in Water Resources [152].

1.2.2 Extracting design information

Rainfall only affects hydropower generation insofar as it affects streamflow distribution. Extracting design information (streamflow distribution) from available data (daily rainfall) is the second challenge faced by local communities. While simple empirical and statistical approaches exist to model the temporal distribution of streamflow, these models fail to properly account for the complexity of streamflow generating processes. They require extensive calibration, have little external validity and cannot cope with non stationary conditions induced by climate change. I take an alternate, process-based probabilistic approach to model streamflow duration. Based on simple assumptions about both the stochastic structure of rainfall and the watershed response, I derive an analytical expression for streamflow distribution that accounts for the strong seasonality of precipitation in Nepal and allows disentangling intra-annual and inter-annual streamflow variations. The model only requires a small number of physically meaningful and observable parameters and has minimal calibration requirement. The method is derived and successfully validated in Nepal, Western Australia and coastal California in Chapter 3, which was published as a research article (co-authored with Sally Thompson and David Dralle) in Water Resources Research [153].

1.2.3 Getting locally relevant information

Unlike surface water processes (e.g., precipitation, evapotranspiration), groundwater processes cannot directly be observed from space-borne platforms. The corresponding parameters of the streamflow distribution model (in particular the recession constant) must therefore be calibrated using discharge observations at available gauges, and spatially interpolated to the catchment and community of interest. Unfortunately, standard geostatistical techniques (e.g., kriging) are not applicable because the structure of the stream network affects the spatial distribution of streamflow variables. Existing approaches to account for network topology are computationally intensive and provide a biased estimate of prediction uncertainty. I developed a new geostatistical approach to streamflow regionalization that incorporates network topology and observable covariates. The new method significantly improves computational performance, the prediction of streamflow variables and the estimation of prediction uncertainties, as shown in the comparative cross validation analysis in Nepal and Austria reported in Chapter 4. The content of the chapter was published as a discussion paper (co-authored with Sally Thompson) in Hydrology and Earth System Sciences Discussions [154] and is in open review for Hydrology and Earth System Sciences.

1.2.4 Inferring economic behavior

In contrast to grid-connected infrastructure, where excess power can be transmitted to demand centers, the feasibility of isolated micro-hydropower schemes is affected by local electricity demand. In particular, the price elasticity of demand will determine the range of infrastructure capacities (i.e. the size of the turbine) that will allow cost recovery. Unfortunately, the experimental setup necessary to rigorously estimate price elasticity[7] is not feasible because households do not have access to electricity in the first place. For a similar reason, any household survey administered by project managers on their community is amenable to significant hypothetical biases. In Chapter 5, I use an openly available cross-sectional infrastructure survey in an instrumental variable setup to estimate the price elasticities of household electricity demand and micro-hydro power supply. While the identification of relevant and valid instruments is a recurring challenge in the development economics literature, I propose a two-instruments approach using topography and census-based variables to simultaneously identify local supply and demand curves for micro-hydropower. The resulting estimates are in line with previously estimated elasticities and highlight important tradeoffs between behavioral accuracy and empirical tractability.

1.2.5 Optimizing and evaluating infrastructure

The last barrier relates to the ability to use the previously obtained design inputs to generate and evaluate infrastructure variants. While significant research has been devoted to optimizing the placement of hydropower infrastructure using digital elevation models, few studies consider run-of-river schemes and none (to my knowledge) incorporates the slope of hillsides in the optimization[8]. The tool presented in Chapter 6 explicitly considers local topography to provide optimized locations for diversions canal and penstocks at the most topographically suitable sites within the considered community. The tool also integrates routines to use design inputs from Chapters 2 to 5 to optimize the size of the infrastructure and evaluates its economic performance.

Lastly, currently available open source softwares allow a seamless integration between geographic information systems, statistical and numerical engines and high level web frameworks. Also described in Chapter 6 is a combination of these platforms that integrates this dissertation's models in a web-based graphical user interface with an interactive map. In addition to enabling smartphone-equipped community members to evaluate the feasibility of micro-hydropower in their own community, the tool stores individual community assessment in a geographic database and generate large-scale feasibility maps of micro hydropower in Nepal.

[7] Price elasticity measures the effect of price changes on household electricity consumption. Its rigorous identification requires a randomized experiment, where households are imposed a randomly selected change in the unit price of their consumed electricity.

[8] Hillsides being typically steeper than stream beds, micro-hydropower plants are usually located at a significant distance from the corresponding stream intake. Diversions up to several kilometers have been observed in Nepal.

Chapter 2

Bias adjustment of satellite rainfall data through stochastic modeling: Methods development and application to Nepal

2.1 Introduction

Spatially explicit rainfall estimates are crucial for hydrologic predictions, but due to challenges in observing rainfall at watershed scales, rainfall estimates remain a major source of uncertainty for hydrologic models [245]. In many parts of the world, ground-based rain-gauge networks are irregular and locally sparse [126], and may be biased with respect to the sources of environmental variability (see Figure 2.1 for an example). Such networks do not provide a robust basis for inferring the spatial pattern of rainfall fields. An alternative and explicitly spatial rainfall product is provided by satellite observations of precipitation. Unfortunately, satellite observations of rainfall have widely acknowledged limitations, including sensitivity to precipitation type [110], underestimation of orographic rainfall [58], a tendency to miss snowfall [12], inability to capture short rainfall events [33] and systematic biases in mountainous areas [203, 234, 12, 117]. Using ground-based data to correct biases in satellite data provides one method to address these limitations. For example, the satellite observations in the NASA Tropical Rainfall Measuring Mission (TRMM) 3B42 dataset are adjusted using monthly-averaged ground observations provided by local monitoring agencies to the Global Precipitation Climatology Centre (GPCC) [110]. However, the efficiency of the adjustment is limited by the scarcity of available gauges and typically requires careful regional evaluation against local precipitation measurements.

The correction applied by NASA on TRMM is a standard bias adjustment procedure for satellite rainfall observations, based on correcting rainfall time series - in this case by regression analysis applied to cumulative rainfall totals [203, 249, 57]. Other standard pro-

cedures adjust quantiles of the daily rainfall to match those observed at gauges [98]. These approaches suffer from several drawbacks:

1. Biases in TRMM observations of rainfall timeseries are influenced by errors in both rainfall frequency and rainfall intensity, which may have opposite signs [76]. Adjusting satellite precipitation totals or probability density functions (PDF) will thus correct errors in the magnitude of rainfall, but not in its temporal structure, although both factors are important for hydrological predictions [133, 38].

2. Although some recent studies account for the observed spatial heterogeneity in biases, and in doing so significantly improved the corrected dataset [249, 57], approaches based on preserving regional rainfall totals often do not account for spatial patterns in bias or focus on single precipitation stations. One of the factors that makes spatially-explicit corrections challenging is the upscaling of point observations from gauges to areal rainfall at the resolution of the satellite grid.

3. Finally, correction of monthly time series on a pixel by pixel basis is numerically intensive, and cannot take advantage of historical rainfall datasets which, although not overlapping with contemporary observations, may still contain useful information about spatial patterns in rainfall.

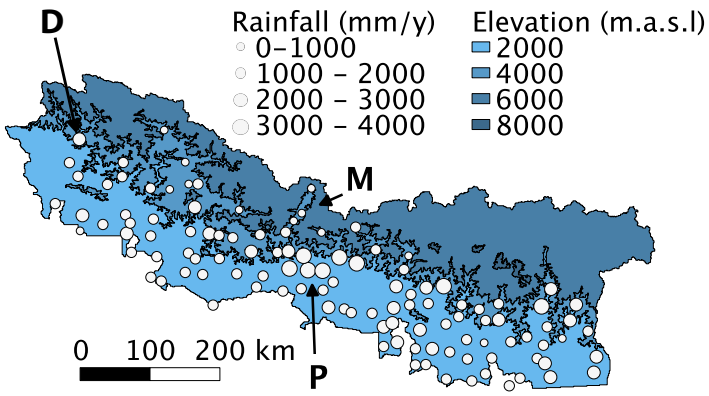
I therefore propose an alternative strategy for bias adjustment of satellite rainfall data using ground-based gauge observations. Instead of adjusting daily rainfall to match the mean monthly precipitation, I perform the bias adjustment on a set of (pseudo)stationary stochastic parameters that describe the rainfall process in terms of frequency, intensity, and the autocorrelation of wet and dry periods [213, 184, 121]. This approach addresses the key limitations of time series based bias adjustment:

1. It is a direct response to the observation of different directionality in TRMM-gauge bias arising due to different and independent features of the rainfall time series [76]. This observation implies that separating the bias adjustment for rainfall occurrence and intensity might improve the robustness of the resulting rainfall estimates.

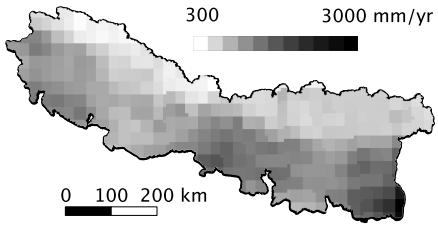
2. It allows different features of rainfall to be independently interpolated, accounting for spatial heterogeneity and, unlike existing studies, also accounting for potential differences in spatial heterogeneities between stochastic rainfall features.

3. Being in the frequency domain, the bias adjustment can be operated using non overlapping observed time series provided stationarity conditions are satisfied.

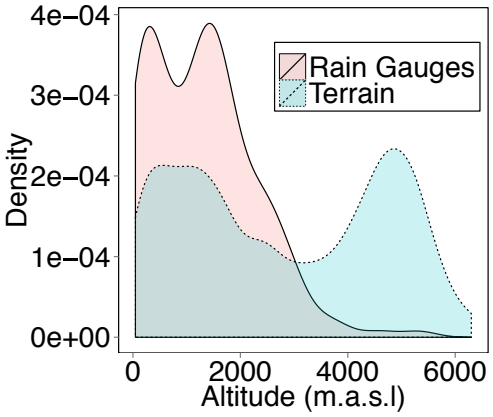
A key contribution of the proposed procedure lies in its ability to spatially aggregate and interpolate the stochastic rainfall descriptors at the grid resolution. This provides a ground



(a)



(b)



(c)

Figure 2.1: Study region and available data. (a) Location of the available gauges and mean annual rainfall. The figure shows vast zones in the North that are not covered by the gauge network. The difference in annual rainfall between Pokhara (P) and Mustang (M), two proximate regions separated by the Annapurna range, illustrates the importance of rain shadow effects. The example of time series correction described in Part (f) focuses on the rainfall gauge at Darchula (D) in western Nepal. (b) Yearly rainfall in 2010 measured by TRMM 3B43 v6 (monthly precipitation) and aggregated annually, showing decreasing trends towards the east and north. (c) Kernel density estimates of the altitude distributions of the area and of the rain gauges. The figure shows that the altitude distribution of the area is bimodal with modes at 1000 masl and 5000 masl. This distribution is not matched by the gauges, which are preferentially located below 3000 masl.

truth estimate of the daily rainfall distribution at each pixel that can be used to correct satellite rainfall distributions, with two potential applications. Firstly, grid-scale rainfall cumulative probability densities (CDF) are valuable for correcting rainfall timeseries magnitudes via quantile mapping [98]. My proposed method explores the upscaling of gauge-derived rainfall PDFs and their spatial interpolation, allowing corrections to the rainfall CDF to be applied in a spatially explicit fashion. Moreover, the procedure upscales and interpolates information about the autocorrelation of rainfall, allowing the bias adjustment procedure to correct the temporal structure of satellite rainfall observations as well as the magnitudes. Since the temporal structure of rainfall is an important driver of hydrological responses in the vadose zone [133] and in the flow regime [38], incorporating this information into satellite bias correction is a useful advance. The stochastic parameters may be directly utilized in a stochastic description of the resulting streamflow as presented in Chapter 3; used to generate ensembles of synthetic time series data using stochastic weather generation models [237, 43], or incorporated into time-series correction approaches (as outlined in Section 2.2.5).

The proposed approaches are illustrated here using Nepal as a case study. Nepal provides an excellent opportunity to test the new bias correction procedure because two satellite rainfall products are available that incorporate very different bias-adjustment techniques: TRMMv6 and TRMMv7. The major distinction between the two datasets for terrestrial rainfall estimates lies in the rain gauge datasets used for monthly bias adjustment [109]. In Nepal the number of considered gauges increases from 11 (GPCC monitoring dataset v2) to 280 gauges (GPCC full analysis dataset v6). Thus, TRMMv6 in Nepal represents a satellite rainfall data product with minimal ground-based correction, while TRMM v7 represents satellite data corrected using conventional time series adjustment. In this study, I therefore develop a bias adjustment technique, apply it to TRMM v6 and compare the results against the performance of TRMM v7 as a benchmark.

I first describe a stochastic rainfall model (Section 2.2.1) and its use to adjust satellite rainfall observation biases through space. Spatial adjustment of stochastic parameters is not straightforward because of their nonlinear relationships to the moments and time-structure of the rainfall distribution. To estimate bias, the stochastic model parameters obtained from point-scale rainfall measurements at gauges are spatially aggregated to the scale of a satellite observation pixel (Section 2.2.2). The stochastic model parameters estimated at the pixel scale are then spatially interpolated to provide estimates at the satellite pixels devoid of gauges (Section 2.2.3). Section 2.2.4 summarizes the method to correct the bias of gridded, remotely sensed daily rainfall observations in the frequency domain using multi-site gauge observations – the main contribution of this paper. Using bias adjusted frequency domain information, rainfall time series can then easily be adjusted through quantile mapping (Section 2.2.5). An illustrative example of time series correction is given in Part (f). The remainder of the paper focuses on assessing the performance of the frequency domain bias correction method, which underpins both the stochastic and time-series adjustments. The sensitivity of the method to common sources of uncertainties is first assessed in a Monte Carlo analysis (Section 2.3.1), and its ability to adjust the frequency, mean intensity and variance of actual remote sensing rainfall data is assessed in a cross validation analysis using

Nepalese rainfall for various densities of gauge networks (Section 2.3.2). The main results and their implications are discussed in Section 2.4 and Section 2.5 concludes.

2.2 Theory

2.2.1 Stochastic Model

I use a two-step stochastic weather generator to represent the statistical properties of the rainfall time series. I firstly disaggregate the time series into two independent seasons [213] – the dry season and the monsoon. I identify the seasons by the calendar days corresponding to the average start date (RnStr) and end date (RnStp) of the monsoon. Next, I describe the rainfall for each season in terms of two stochastic processes: the daily occurrence, and daily intensity of rainfall. I use a first-order Markov chain model to represent rainfall occurrence [184, 121]. This model is governed by two parameters P_{01} and P_{11}, which characterize the probability of a rainy day, conditional on the previous day being dry (P_{01}) or rainy (P_{11}). I use a gamma distribution with shape parameter GS and rate parameter GR to describe the probability distribution of daily rainfall depths on those days when rain occurred. This representation of rainfall requires a total of 10 *stochastic model parameters* (SMPs) listed in Table 2.2. These model parameters are directly related to a range of relevant metrics that describe rainfall distribution and can thus be used to evaluate the bias adjustment method. These metrics are derived in Appendix A.1 and include the length of wet and dry spells, the number of rainy days per year, the unconditional variance on daily rainfall and the average annual rainfall .

2.2.2 Areal Aggregation of Stochastic Model Parameters

While gauges monitor precipitation at particular points, satellites observe an areally averaged value of rainfall over many square kilometers. Correcting remote sensing precipitation observations therefore requires spatially aggregating point-scale precipitation parameters to the level of the satellite resolution. I perform this aggregation analytically, rather than directly from the time series because (i) it is more computationally efficient and (ii) it allows us to use data provided by on (stationary) rainfall gauges that do not overlap in time with the TRMM observation window. I outline the applicability of the methods to the case study with TRMM in Nepal below, including an evaluation of the stationarity of ground-based rainfall measurements in terms of the 10 SMPs.

(a) Seasonal Parameters

I assume that the starting day of the rainy and dry seasons at the pixel level can be approximated by the weighted average of the corresponding values across the N_p gauges in the

SMP	Stochastic Model Parameter
MAE	Mean Absolute Error
X	Daily Precipitation
P	Probability of daily precipitation occurrence
N_p	Number of gauges in the considered pixel
N_g	Number of gauges to interpolate from
N_{MC}	Number of Monte Carlo rounds
N_{CV}	Number of Cross Validation rounds
a_i	Fraction of pixel occupied by the Thiessen polygon of gauge i
$v_i^{(\mathsf{param})}$	Interpolation weight associated to gauge i and parameter param
$C(d)$	Correction factor on variance for areal rainfall on in a pixel of diagonal d
$_i$	Subscript for gauges (point rainfall)
$_{\mathsf{pt}}$	Subscript for pixels aggregated from gauges (point rainfall)
$_{\mathsf{pix}}$	Subscript for pixels aggregated from gauges (areal rainfall)
$_{\mathsf{TRMM}}$	Subscript for TRMM pixels
$_{\mathsf{wet}}$	Subscript for rainy days
$_j$	Subscript for Monte Carlo or Cross Validation rounds

Table 2.1: Acronyms, variables and subscripts

$P_{01}^{(w)}$	Probability of a dry day being followed by a wet day in the wet season
$P_{01}^{(d)}$	Probability of a dry day being followed by wet day in the dry season
$P_{11}^{(w)}$	Probability of a wet day being followed by wet day in the wet season
$P_{11}^{(d)}$	Probability of a dry day being followed by wet day in the wet season
$\mathsf{GS}^{(w)}$	Gamma shape parameter for daily rainfall depth in the wet season
$\mathsf{GS}^{(d)}$	Gamma shape parameter for daily rainfall depth in the dry season
$\mathsf{GR}^{(w)}$	Gamma rate parameter for daily rainfall depth in the wet season
$\mathsf{GR}^{(d)}$	Gamma rate parameter for daily rainfall depth in the dry season
RnStr	Average calendar day when Monsoon starts
RnStp	Average calendar day when Monsoon ends

Table 2.2: Stochastic Model Parameters (SMP).

pixel,

$$\mathsf{Str}_{\mathsf{pix}} = \sum_{i=1}^{N_p} a_i \mathsf{Str}_i \qquad \mathsf{Str} \in \{\mathsf{RnStr}, \mathsf{RnStp}\} \tag{2.1}$$

where a_i is the proportion of the pixel's area covered by a Thiessen polygon centered on gauge i.

(b) Occurrence Parameters

A pixel should be classified as 'rainy' on a given day if rain occurs at *any* of its gauges during that day. Thus the probability of rain at the level of a pixel is not a simple average of the occurrence probabilities at the gauges within the pixel, but is modified by the correlation between the gauges. If the correlation length-scale of rainfall exceeds the pixel size, then it is reasonable to assume that the correlation between the rain occurrence probabilities P_i at the different gauges is positive and maximal. That is, if the gauge that is most likely to receive rainfall is dry, the pixel is also dry. Using this assumption, the probability of rainfall in a pixel is well approximated by the maximum occurrence probability across the N_p gauges within that pixel, as:

$$P_{\mathsf{pix}} \approx \max P_i \tag{2.2}$$

A similar assumption about the ratio of wet-to-wet transitions $P_i \cdot P_{11,i}$ leads to the following estimate for the pixel-level transition probability:

$$P_{11,\mathsf{pix}} \approx \max \left\{ \frac{\max\left(P_i \cdot P_{11,i}\right)}{P_{\mathsf{pix}}}; 1 - \frac{\sum_{i=1}^{N} P_i \cdot (1 - P_{11,i})}{P_{\mathsf{pix}}} \right\} \tag{2.3}$$

where the transition probability P_{11} at the satellite pixel level can be approximated by its lower bound. This bound is given by the higher of (i) the maximal value of wet-to-wet ratio ($P \cdot P_{11}$) and (ii) the sum of wet-to-dry transition ratios ($P \cdot P_{10}$) within that pixel. The full derivation of equations Equation (2.2) and Equation (2.3) is presented in Appendix A.2. Our case study in Nepal is characterized by a maximum density of 5 gauges per pixel and spatial autocorrelation ranges of approximately 3 (dry season) to 4 (wet season) times the pixel size of 27.7km (Table 2.4), meeting the assumptions used in the derivation of equations Equation (2.2) and Equation (2.3). I tested the performance of the aggregation equations via a Monte Carlo analysis. I found that using equations Equation (2.2) and Equation (2.3) generated less than 2% error in both metrics (P_i and $P_i \cdot P_{11,i}$). This error declined with an increase in the correlation length scale, but increased with increasing numbers of gauges per pixel.

(c) Intensity Parameters

To aggregate rainfall intensity I preserve the weighted average of the first two moments of the distributions measured at each gauge, using the Thiessen polygon area ratios a_i as

weights. Doing so based on the SMPs that describe the rainfall intensity (GS and GR) poses three challenges. Firstly, the SMPs are non linearly related to the moments of the gamma distribution:

$$\mathsf{E}\left[X \mid \mathsf{wet}\right] = \mathsf{GS}/\mathsf{GR} \tag{2.4}$$

$$\mathsf{Var}\left(X \mid \mathsf{wet}\right) = \mathsf{GS}/\mathsf{GR}^2. \tag{2.5}$$

Thus, aggregating the weighted sum of the distribution's parameters is not equivalent to aggregating the distribution's moments. Secondly, the parameters represent the distribution of rainfall intensity *conditional* on rainfall occurrence, so the probability P of rainfall occurrence must be incorporated into the aggregation. Finally, the variance of areal rainfall is affected by spatial autocorrelation. A full derivation of the upscaling relationship for the rainfall intensity properties, accounting for these three challenges, is provided in Appendix A.3. The methodology used consists of (i) conditioning for rainfall occurrence and the location of individual gauges, (ii) applying the laws of iterated expectation and total variance to compute the mean and variance of rainfall intensity at the pixel scale (Equation (2.6) and Equation (2.7)) and (iii) correcting the variance of areal rainfall to account for the transition from point to areal probabilities [188]. I assume the same functional form of the PDF applies to pixels and all gauges, meaning that the pixel-scale rainfall intensity is a gamma distribution and that its parameters GS and GR are directly related to its mean and variance as in Equation (2.4) and Equation (2.5). With these assumptions, I obtain the expectation and variance of the pixel-level areal rainfall as:

$$\mathsf{E}\left[X_{\mathsf{pix}} \mid \mathsf{wet}\right] = \frac{1}{P_{\mathsf{pix}}} \cdot \sum_{i=1}^{N} a_i P_i \mathsf{E}\left[X_i | \mathsf{wet}\right] \tag{2.6}$$

$$\begin{aligned}\mathsf{Var}\left(X_{\mathsf{pix}} \mid \mathsf{wet}\right) &= \\ &= \frac{C(d)}{P_{\mathsf{pix}}}\left[\sum_{i=1}^{N} a_i P_i \left(\mathsf{Var}\left(X_i \mid \mathsf{wet}\right) + P_i \mathsf{E}\left[X_i \mid \mathsf{wet}\right]^2 - P_i \mathsf{E}\left[X_i \mid \mathsf{wet}\right]\right)\right] \\ &+ C(d) P_{\mathsf{pix}}\left[\mathsf{E}\left[X_{\mathsf{pix}} \mid \mathsf{wet}\right] - \mathsf{E}\left[X_{\mathsf{pix}} \mid \mathsf{wet}\right]^2\right],\end{aligned} \tag{2.7}$$

where P_i is the probability of rainfall occurrence at the gauge level, and P_{pix} is the probability of rainfall occurrence at the pixel level (from Equation (2.2)). $C(d)$ is an attenuation factor applied to the variance of areal rainfall based on the derivation of Rodriguez-Iturbe and Mejía [188]:

$$C(d) = \int^{\sqrt{2}d} r(\nu) f(\nu) \mathrm{d}\nu \leq 1,$$

where $r(\nu)$ is the spatial correlation function of rainfall intensity and $f(\nu)$ is the distribution of distances between two points chosen at random in the pixel. Point-scale rainfall typically over-estimates the variance of areal rainfall, so $C(d) < 1$. $C(d)$ increases with pixel size d

and decreases with the spatial autocorrelation range, both of which are typically spatially homogenous. In Nepal I estimated $C(27.7\mathsf{km})$ as 0.75 in the monsoon and 0.86 in the dry season, using a correlogram estimated from the spatial distribution of rainfall intensity at gauges over 2,000 randomly selected days.

2.2.3 Spatial Interpolation of Stochastic Model Parameters

A typical spatial interpolation methodology would approximate daily rainfall $\tilde{X}$ at unmonitored locations as linear combinations of X_i the rainfall measured at surrounding locations i on the same day, weighted by v_i^X, a normalized similarity metric based on relative position (e.g. inverse weighted distance) or the spatial correlation function of X (e.g. kriging):

$$\tilde{X} = \sum_{i=1}^{N_g} v_i^{(X)} X_i \tag{2.8}$$

Interpolation of the probabilistic descriptors of the rainfall, however, cannot be undertaken by directly interpolating the SMP's because neither the moments of the gamma distribution of conditional rainfall intensity nor the moments of the binomial distribution of daily rainfall occurrence are linear combinations of the SMPs. Thus, I interpolate the moments of the distributions, expressed as functions of the SMPs. I assume that interpolation must preserve seasonal transition dates (RnStr and RnStp), the daily occurrence probability of rainfall (P) and the ratio of wet-to-wet transitions ($P \cdot P_{11}$). This allows us to express the interpolated rainfall metrics as linear combinations of their respective values at the N_g observed locations, which are directly related to the observed SMPs:

$$\widetilde{\mathsf{Str}} = \sum_{i=1}^{N_g} v_i^{(\mathsf{Str})} \mathsf{Str}_i \qquad \mathsf{Str} \in \{\mathsf{RnStr}, \mathsf{RnStp}\} \tag{2.9}$$

$$\tilde{P} = \sum_{i=1}^{N_g} v_i^{(P)} P_i = \sum_{i=1}^{N_g} v_i^{(P)} \frac{P_{01,i}}{1 + P_{01,i} - P_{11,i}} \tag{2.10}$$

$$\widetilde{P \cdot P_{11}} = \sum_{i=1}^{N_g} v_i^{(P \cdot P_{11})} P_i \cdot P_{11,i} = \sum_{i=1}^{N_g} v_i^{(P \cdot P_{11})} \frac{P_{11,i} P_{01,i}}{1 + P_{01,i} - P_{11,i}} \tag{2.11}$$

Using similar reasoning to that in Section 2.2.2, but replacing area weights a_i with interpolation weights $v_i^{(\mathsf{E})}$, I compute the interpolated moments of the distribution of conditional rainfall intensity. Here I use weights $v_i^{(\mathsf{E})}$ generated from kriging of the expected rainfall $\mathsf{E}[X_i]$ for the interpolation of both the mean and variance of the rainfall PDF. Either ordinary kriging or univerval kriging can be used [171, 183]. For this interpolation, I do not use the attenuation factor $C(d)$ as there is no point to area transformation. From equations Equation (2.6) and Equation (2.7) I obtain the expectation and variance of the rainfall at

the ungauged location:

$$\mathsf{E}\left[\tilde{X} \mid \mathsf{wet}\right] = \frac{1}{\tilde{P}} \cdot \sum_{i=1}^{N_g} v_i^{(\mathsf{E})} P_i \mathsf{E}\left[X_i | \mathsf{wet}\right] \tag{2.12}$$

$$\begin{aligned}\mathsf{Var}\left(\tilde{X} \mid \mathsf{wet}\right) &= \\ &= \frac{1}{\tilde{P}} \left[\sum_{i=1}^{N_g} v_i^{(\mathsf{E})} P_i \left(\mathsf{Var}\left(X_i \mid \mathsf{wet}\right) + P_i \mathsf{E}\left[X_i \mid \mathsf{wet}\right]^2 - P_i \mathsf{E}\left[X_i \mid \mathsf{wet}\right]\right)\right] \\ &+ \tilde{P}\left[\mathsf{E}\left[\tilde{X} \mid \mathsf{wet}\right] - \mathsf{E}\left[\tilde{X} \mid \mathsf{wet}\right]^2\right]\end{aligned} \tag{2.13}$$

where P_i is the probability of rainfall occurrence at the observation point i, and $\tilde{P}$ the interpolated probability of rainfall given by Equation (2.10).

2.2.4 Bias adjustment of Stochastic Model Parameters

The bias adjustment approach is based on the assumption of spatial correlation in the differences in daily rainfall between the TRMM pixels and the (aggregated) gauges. Biases at pixels devoid of gauges can then be estimated by interpolating the biases observed at pixels that contain gauges. Interpolating the biases for each stochastic parameter to un-gauged pixels raises the same problems as interpolating the stochastic parameters within the pixels (Section 2.2.3). Thus, I independently interpolate the SMPs estimated from TRMMv6 at gauged pixels and the pixel-scale SMPs estimated from the gauges (and *not* the difference between them), before computing the biases at ungauged pixels as the difference between the two interpolations. The full bias adjustment procedure thus consists of the following steps:

(i) Aggregating the SMPs observed at the gauges to the resolution of TRMM pixels (Section 2.2.2).

(ii) Interpolating the aggregated SMPs from the gauged to the ungauged pixels (Section 2.2.3), labeled as $\tilde{\mathsf{SMP}}_{\mathsf{pix}}$.

(iii) Interpolating the SMPs obtained for TRMMv6 at the gauged pixels to the ungauged pixels (Section 2.2.3), labeled as $\tilde{\mathsf{SMP}}_{\mathsf{TRMM}}$.

(iv) Computing the biases $\Delta\tilde{\mathsf{SMP}}_{\mathsf{TRMM}}$ at ungauged pixels by subtracting the result of step (ii) ($\tilde{\mathsf{SMP}}_{\mathsf{pix}}$) to the result of step (iii) ($\tilde{\mathsf{SMP}}_{\mathsf{TRMM}}$).

(v) Finally, biases are adjusted by subtracting the modeled bias $\Delta\tilde{\mathsf{SMP}}_{\mathsf{TRMM}}$ from $\mathsf{SMP}_{\mathsf{TRMM}}$, the local SMPs of TRMMv6:

$$\mathsf{SMP}_{\mathsf{adjusted}} = \mathsf{SMP}_{\mathsf{TRMM}} - \Delta\tilde{\mathsf{SMP}}_{\mathsf{TRMM}} = \mathsf{SMP}_{\mathsf{TRMM}} - \left(\tilde{\mathsf{SMP}}_{\mathsf{TRMM}} - \tilde{\mathsf{SMP}}_{\mathsf{pix}}\right).$$

Assuming rainfall follows the stochastic model described in Section 2.2.1, this procedure allows the bias adjusted distribution of rainfall to be estimated for all pixels.

2.2.5 Bias adjustment of time series

A useful application of the bias adjusted distribution of rainfall obtained in the previous is its use to correct remotely sensed time series through quantile mapping. Quantile mapping is a well established technique (see [98] for a review) that, in the context of this paper, attempts to find a transformation of $X^{(t)}_{TRMM}$, the remotely sensed rainfall observation at time t, such that its new distribution equals the distribution of $X^{(t)}_{\mathsf{adj}}$, the corresponding bias adjusted rainfall observation. The distribution of $X^{(t)}_{TRMM}$ can be readily characterized from remote sensing observations. The method presented in Part (c) provides the bias corrected distribution of rainfall (i.e the distribution of $X^{(t)}_{\mathsf{adj}}$). The transformation can therefore be written as

$$X^{(t)}_{\mathsf{adj}} = F^{-1}_{\mathsf{adj}}(F_{\mathsf{TRMM}}(X^{(t)}_{\mathsf{TRMM}})) \tag{2.14}$$

where $F^{-1}_{\mathsf{adj}}(\cdot)$ is the inverse of the bias adjusted cumulative distribution function and $F_{\mathsf{TRMM}}(\cdot)$ is the cumulative distribution function of remotely sensed rainfall at the considered pixel. $F_{\mathsf{TRMM}}(X^{(t)})$ can be calculated using the relevant stochastic model parameters obtained from remotely sensed rainfall by applying the law of total probabilities:

$$F_{\mathsf{TRMM}}(X^{(t)}_{\mathsf{TRMM}}) = (1 - P_{\mathsf{TRMM}}) + P_{\mathsf{TRMM}} \cdot F_{\mathsf{TRMM,w}}(X^{(t)}_{\mathsf{TRMM}}) \tag{2.15}$$

where $P_{\mathsf{TRMM}} = P_{01,\mathsf{TRMM}}$ if $X^{(t-1)}_{\mathsf{TRMM}} = 0$ and $P_{\mathsf{TRMM}} = P_{11,\mathsf{TRMM}}$ otherwise; and where $F_{\mathsf{TRMM,w}}(X^{(t)})$ is the cumulative distribution function of a gamma distribution with rate GR_{TRMM} and shape GS_{TRMM}. Similarly, the bias-adjusted cdf F_{adj} can be calculated using the bias-adjusted stochastic model parameters.

$$F_{\mathsf{adj}}(X^{(t)}_{\mathsf{adj}}) = (1 - P_{\mathsf{adj}}) + P_{\mathsf{adj}} \cdot F_{\mathsf{adj,w}}(X^{(t)}_{\mathsf{adj}}) \tag{2.16}$$

where $P_{\mathsf{adj}} = P_{01,\mathsf{adj}}$ if $X^{(t-1)}_{\mathsf{adj}} = 0$ and $P_{\mathsf{adj}} = P_{11,\mathsf{adj}}$ otherwise; and where $F_{\mathsf{adj,w}}(Y^{(t)})$ is the cumulative distribution function of a gamma distribution with rate $\mathsf{GR}_{\mathsf{adj}}$ and shape $\mathsf{GS}_{\mathsf{adj}}$. I define the inverse of $F_{\mathsf{adj}}(\cdot)$ as

$$F^{-1}_{\mathsf{adj}}(Y^{(t)}) = \begin{cases} 0 & \text{if } Y^{(t)} \leq 1 - P_{\mathsf{adj}} \\ F^{-1}_{\mathsf{adj,w}}(Y^{(t)}) & \text{otherwise} \end{cases} \tag{2.17}$$

Note that $F_{\mathsf{TRMM}}(\cdot)$ has a discontinuity at zero. Therefore, its image doesn't span all possible probabilities between zero and one (i.e. values below P_{TRMM} are excluded from the image). When applying quantile mapping (Equation (2.14)) part of the rainfall range is therefore censored. For example if $F_{\mathsf{adj}}(0) < 1 - P_{\mathsf{TRMM}}$, all values of $X^{(t)}_{\mathsf{TRMM}}$ will be mapped to positive rainfall.[1] In other words, a dry data point in TRMM is always matched to the *largest* rainfall value $X^{(t)}_{\mathsf{adj}}$ that occurs with the probability $F_{\mathsf{TRMM}}(0)$ in my model.

[1] One particular concern is artificial oscillation of rainfall occurrence during dry periods, when $P_{adj,01} < P_{\mathsf{TRMM}} < P_{adj,11}$ (or $P_{adj,11} < P_{\mathsf{TRMM}} < P_{adj,01}$).

Of course, any rainfall prediction below this cutoff would be just as reasonable. To avoid artificial overestimation of rainfall occurrence, I therefore match a dry TRMM data point to a random sample from the conditional distribution $F_{\mathsf{adj}}(x \mid x \leq X_{\mathsf{adj}}^{(t)})$, given by

$$F_{\mathsf{adj}}(x \mid x \leq X_{\mathsf{adj}}^{(t)}) = \begin{cases} \frac{F_{\mathsf{adj}}(x)}{F_{\mathsf{adj}}(X_{\mathsf{adj}}^{(t)})} & \text{if } x \in [0, X_{\mathsf{adj}}^{(t)}] \\ 1 & \text{if } x > X_{\mathsf{adj}}^{(t)}. \end{cases} \tag{2.18}$$

This correction ensures that I preserve the actual rainfall distribution (including rainfall occurrence) for large samples.

To summarize, I first determine which stochastic model parameters to use according to the season of $X_{TRMM}^{(t)}$ (Monsoon vs. dry season) and the rainfall occurrence status at $X_{TRMM}^{(t-1)}$ (wet vs. dry). Then,

- if $X_{TRMM}^{(t)} > 0$, I apply Equation (2.15) to get the probability of $X_{TRMM}^{(t)}$, on which I finally apply Equation (2.17) to get the corresponding quantile in the adjusted rainfall distribution.
- if $X_{TRMM}^{(t)} = 0$ I have $F_{\mathsf{TRMM}}(X_{\mathsf{TRMM}}^{(t)}) = P_{\mathsf{TRMM}}$ and are confronted to the discontinuity problem mentioned above. The case where $F_{\mathsf{adj}}^{-1}(P_{\mathsf{TRMM}}) = 0$ results in a dry day and $X_{\mathsf{adj}}^{(t)} = 0$. If $F_{\mathsf{adj}}^{-1}(P_{\mathsf{TRMM}}) > 0$, $X_{\mathsf{adj}}^{(t)}$ is stochastically determined as a random draw from the distribution, which cdf is described in Equation (2.18). This is equivalent to the practically more convenient option of a random draw from the distribution in Equation (2.16) with rejection of samples above $F_{\mathsf{adj}}^{-1}(P_{\mathsf{TRMM}})$.

2.3 Methods

The methods describes the metrics used to evaluate the bias adjustment process of stochastic model parameters described in Section 2.2.4, and a Monte Carlo analysis in which the performance of the process was tested on synthetic data It then outlines the application of the technique to rainfall data in Nepal. As part of this application I characterize the bias in TRMM observations , and perform a jack-knife cross validation [201] to assess the performance of the bias-adjustment technique. An example of the application of the adjusted stochastic model parameters to correct TRMM time series is finally proposed to close this section. The stochastic model, bias adjustment methods and time series correction procedure were compiled in an R script [113] and are provided as supplementary material.

2.3.1 Monte Carlo Analysis

To evaluate the performance of the bias adjustment I focus on the mean absolute errors (MAE) in annual rainfall. The MAE avoids outlier compensation effects, whereby overestimation at one gauge may cancel out the underestimation at another (leading to underestimation of the true error). The MAE of annual rainfall provides a scalar performance

metric that combines errors in the occurrence, intensity and seasonality of rainfall and is easily understood in physical terms. I also compute MAEs for the variance and occurrence probability of daily rainfall.

I run a Monte Carlo analysis using synthetic data to evaluate the properties of my bias adjustment technique and its sensitivity to a range of characteristics of the gauge network and TRMM observations (presented in Table 2.3).

Variable	Default Value	Experimental Range
fAWN_{local}	0	0 - 0.2
N	50	10 - 1000
z_{max}	8848	1000 - 8848
fAWN_{obs}	0	0 - 0.3
fBIAS_{mean}	1	0.5 - 5
fBIAS_{range}	1	0.01 - 2

Table 2.3: Experimental variables, their default value and range considered in the Monte Carlo experiments. fAWN_{local} and fAWN_{obs} represent the standard deviation of local rainfall variations and observation errors on gauges respectively; N and z_{max} represent the size and upper altitude limit of the gauge network; fBIAS_{mean} and fBIAS_{range} the multiplication factors respectively applied on the mean amplitude and spatial autocorrelation range of the TRMM biases observed in Nepal

I apply the following procedure to generate a synthetic rainfall surface, TRMM data and gauge observations that are representative of my case study site (Nepal):

1. The SMP values observed at Nepalese gauges are interpolated by ordinary kriging onto a 0.05° grid, which is generated from a high resolution digital elevation model of Nepal [225].

2. Synthetic SMP surfaces are created by adding white noise (with standard deviation fAWN_{local}) to each point of the grid. This additive noise represents inaccuracies associated with the interpolation and local rainfall variations that are not captured by the gauge network.

3. N grid points are randomly selected as 'rain gauge' locations. I control bias in the selection of gauge locations by specifying an elevation threshold z_{max}, and forcing all gauges to be located below this threshold.

4. Random observation errors are simulated by adding white noise with standard deviation fAWN_{obs} to the SMPs at the synthetic gauges.

5. Synthetic TRMM data are generated by spatially aggregating (Section 2.2.2) the synthetic SMP surfaces at the TRMM resolution of 0.25° and adding a spatially correlated random bias. The mean value and spatial correlation range of the bias are prescribed

as multiples of the corresponding values observed in Nepal with multiplication factors fBIAS_{mean} and fBIAS_{range}.

6. For each of the 'real', bias-corrected and the two control procedures (interpolation of gauges only, or direct use of TRMM observations only), I also generate a surface of the expected annual rainfall, which is used as a basis for computing MAE and evaluating the bias correction technique.

I generate approximately 80 realizations of potential rainfall surfaces by varying each of the parameters in Table 2.3 while maintaining others at the default values listed in Table 2.3. I assess the MAE on the annual rainfall in each case. For each set of numerical experiments, I repeat the Monte Carlo process until the computed MAE becomes insensitive to the addition of further iterations (i.e. changes by less than 1%). The Monte Carlo estimate of the mean absolute error on yearly rainfall ($\mathsf{MAE_{MC}}$) is estimated for the three regionalization procedures: my bias adjustment method, unadjusted (synthetic) TRMM and interpolated (synthetic) gauges. In order to compare the robustness of each procedure to changes in the uncertainty sources in Table 2.3, I normalized all $\mathsf{MAE_{MC}}$ values by the mean absolute error obtained with the default parameter values (Table 2.3). This analysis compares the robustness of the procedures to uncertainty in the input data, but does not evaluate the absolute quality of the rainfall predictions obtained by each method.

2.3.2 Nepal Case Study

(a) Study Area

I used my proposed bias adjustment technique to correct TRMMv6 using rain gauge data in Nepal. Nepal lies on an escarpment bounded by the Gangetic Plain to the south and the Tibetan Plateau to the north. Its large altitudinal range spans diverse physiographic regions, from tropical lowlands to high Himalayan mountains that contain the headwaters of Asia's major river systems and thus water supply for close to 1.4 billion people [116]. This diversity is reflected in the annual rainfall observed at local gauges, which varies from 200 mm y^{-1} in the Trans-Himalayan semi-arid Mustang region, to 4000 mm y^{-1} 100 km further south near the city of Pokhara, upwind of the Annapurna Range (Figure 2.1). I estimated the average annual rainfall of Nepal as 1750mm y^{-1} via Theissen polygon weighting of gauge observations. Most precipitation occurs during the Asian summer monsoon (June to September), when the Himalayan range intercepts strong easterly winds carrying moist air from the Bay of Bengal [27]. The precipitation declines towards the west, reflecting the monsoon circulation. Orography and rain shadows affect rainfall in the high Himalayas and the Tibetan plateau, causing rainfall to also decline towards the north [126]. These regional rainfall patterns reverse in winter (December-February), when westerly weather systems generate snowfall preferentially in the high mountains in Western Nepal. Figure 2.1 shows the spatial pattern in annual rainfall for 2010 as measured by the TRMM 3B43 (v6) monthly rainfall product aggregated at the annual scale. At smaller scales, orographic effects are significant and affect

both the spatial and temporal distribution of rainfall. Daytime rainfall is abundant on ridges, while rain occurs at night, and in smaller volumes, in the valleys [126].

There have been several evaluations of TRMM rainfall predictions in Nepal. TRMMv6 reliably detects monthly rainfall patterns, large-scale rainfall patterns and heavy rainfall events in the Himalayas [27, 14, 21, 244]. At daily time scales, however, TRMMv6 consistently underestimated rainfall volume along the Himalayan range in Nepal [12, 117], while overestimating it on the Tibetan Plateau [249]. A major revision of TRMM 3B42 (TRMMv7) was released in late 2012. In this revision, satellite observations are adjusted using a much larger density of rainfall gauges [175]. As discussed in the introduction, TRMMv6 in Nepal provides us with a barely-corrected satellite rainfall data product, while TRMM v7 provides a comparison with a more traditional method of bias correction, allowing us to benchmark my process against a state-of-the-art bias-adjusted product. I therefore applied the bias correction techniques to TRMMv6 data, treating TRMMv7 as a validation dataset for comparison.

(b) Data Sources and Pre-Processing

Gauge data from 192 rainfall stations for the 1969-1995 period are available from the "Hindu-Kush Himalayan Flow Regimes from International Experimental and Network Data" (HKH-FRIEND) project's Regional Hydrological Data Centre [105]. I obtained additional data from 47 gauges covering a more recent period (1998-2010) from the Department of Hydrology and Meteorology of Nepal [66]. These gauges are a subset of the 280 gauges used to generate the gridded GPCC dataset on which NASA calibrates TRMMv7.

I remove all years that were missing more than 10 days of data and use double mass plots to remove gauges with inhomogeneous data. Different datasets collected at identical locations are merged, generating a final dataset of 114 gauges, with data spans of at least 10 years. I anticipate that considerable observation error remains in this dataset, due to (at least) the diverse range of technologies and data records used at individual gauges. Figure 2.1 shows the gauge locations. Gauges are scarce at elevations above 2000 masl and in the mountainous regions of northern Nepal (Figure 2.1 and Figure 2.1).

Remote sensing precipitation data are obtained from NASA's TRMM 3B42 v6 and v7 research products [160], and aggregated to provide daily rainfall estimates between 1998 and 2010. The daily timescale exceeds the characteristic duration of single rainfall events [27], allowing us to neglect the internal temporal structure of rainfall events.

I test for stationarity of the rainfall fields in the subset of gauges that spanned the whole 1969-2010 period by estimating the value of each SMP over a moving window of 4 years: about 160 rain events. I regress the estimates of the SMPs against time and tested the statistical significance of the regression coefficient with Student-t tests. For gauges where a statistically significant trend was identified ($p<0.01$), I evaluated its impact on the prediction of the annual rainfall over a period of 12 years, which is the average lag between the end of the gauged record and the beginning of the TRMM datasets. For a trend in the SMP to impact the prediction of rainfall, it should generate errors in the annual rainfall prediction

that are comparable to the error associated with the bias adjustment method (22% over 12 years – Part (b)). The majority of gauges (75%) do not have a significant trend in yearly rainfall at the 99% confidence interval. Most (70%) of the gauges with statistically significant rainfall trends do not generate large enough changes in SMPs to affect the bias correction. SMP changes exceeding 22% arose in only 7% of the gauges, mostly on the SMPs related to conditional rainfall intensity: in these gauges, increases in the rate parameter of the gamma distribution were offset by decreases in the shape parameter, leading to little effect on the expected value of rainfall. Therefore, using SMPs computed in the 1969-2010 window provide a valid point of comparison to the SMPs computed from TRMM in the 1998-2010 period in which the satellite operated.

(c) Stochastic Model Fit

I fit the 10-parameter stochastic model to daily precipitation at each gauge and at each TRMM pixel independently. Chi-squared tests confirm significant differences in the P_{01} and P_{11} transition probabilities, validating the use of a Markov chain model for over 90% of the gauges. Kolmogorov-Smirnov and Anderson-Darling tests indicate that a gamma distribution provides the best representation of conditional daily rainfall intensity during the wet season and is comparable to alternative distributions (exponential and log-normal) during the dry season. The calendar days representing the average start and end date of the monsoon (RnStr and RnStp) were identified by fitting a step function to the precipitation time series (Figure 2.4). Once calibrated, the overall performance of the stochastic model was evaluated in terms of mean absolute error, based on its ability to reproduce yearly rainfall as well as the variance and occurrence probability of daily rainfall from the stochastic model parameters.

(d) Bias Adjustment Performance at Gauged Pixels

I verify that removing the biases on the SMPs improves my estimation of the annual rainfall in pixels containing rain gauges. In these pixels, I (i) aggregate the SMPs observed at the gauges to the pixel scale, (ii) correct the SMPs of TRMMv6 using these aggregated values and (iii) evaluate the mean absolute error in estimated yearly rainfall by comparing the adjusted SMPs to rainfall observed at the gauges. The same set of gauges are used to adjust and evaluate the procedure: this first evaluation estimates the combined effects of adjusting the biases in multiple individual parameters at a point, without assessing the effect of aggregating and regionalizing the adjustment.

(e) Bias Adjustment Performance at Ungauged Pixel

I regionalize the adjustments to ungauged pixels by interpolating the SMPs and their biases. I test for spatial trends by running stepwise multiple regressions of the SMP and their respective biases against (i) elevation (as a surrogate for orographic effects), (ii) latitude (as a surrogate for the east-west rainfall trend I anticipated due to Monsoonal circulation patterns) and (iii) longitude (as a surrogate for the north-south rainfall trend I anticipated

due to rain-shadow effects). The coefficients resulting from the optimal combinations of covariates that minimized the Akaike Information Criterion [228] were either not significantly different from zero at the 95% confidence interval, or orders of magnitude smaller than the intercept, allowing us to use ordinary kriging to interpolate the SMPs. The biases in the SMPs were spatially auto-correlated, with ranges above 50km for the stochastic parameters and above 25km for their biases (Table 2.4).

	Season start			P(rain)			E[rain]|wet day			E[wet spells]		
	$\boldsymbol{E}$	$(\boldsymbol{\sigma})$	*Rge*	$\boldsymbol{E}$	$(\boldsymbol{\sigma})$	*Rge*	$\boldsymbol{E}$	$(\boldsymbol{\sigma})$	*Rge*	$\boldsymbol{E}$	$(\boldsymbol{\sigma})$	*Rge*
	[*Cal.day*]		[*km*]	[−]		[*km*]	$\left[\frac{\text{mm}}{\text{day}}\right]$		[*km*]	[day]		[*km*]
Monsoon												
Gauge	158	(14)	*90*	0.65	(0.17)	*125*	19.42	(6.88)	*89*	5.74	(4.55)	*170*
Bias	−1	(13)	*27*	0.09	(0.13)	*86*	−7.19	(5.12)	*54*	0.31	(4.24)	*179*
Dry Season												
Gauge	261	(10)	*56*	0.16	(0.07)	*86*	11.69	(3.13)	*150*	11.88	(5.20)	*128*
Bias	3	(13)	*49*	0.08	(0.10)	*96*	−6.98	(2.40)	*60*	−4.79	(4.51)	*124*

Table 2.4: Seasonal rainfall characteristics in Nepal and related biases. For each season, columns present the calendar day of season start, the probability of rain, the expected rain on a rainy day and the expected length of wet (Monsoon) or dry (dry season) spells. For each parameter, the expected value (E) across the gauge dataset, the standard deviation (σ) and the spatial correlation range (*Rge*) are given .

The performance of the bias adjustment method at ungauged locations is assessed by comparing its performance to the two control methods used in the Monte Carlo analysis: (i) the interpolation of rain gauges and (ii) the direct use of unadjusted TRMMv6. The predictive performance of these three methods is assessed using two independent validation datasets. (i) TRMMv7, which provides an external validation set, and (ii) jack-knife resampling of the ground gauge data, which provides an internal validation set [201]. The jack-knife procedure was applied to predict the pixel-scale rainfall characteristics for twenty percent of the 95 pixels containing rain gauges. A fraction of the remaining gauges was randomly assigned to a training set and used as input for interpolation and bias adjustment. I repeated the jack-knife resampling process approximately 50 times, again terminating the process when adding another replicate caused a change of less than 1% in the MAE. I finally computed the jack-knife estimate of the mean absolute error:

$$\mathsf{MAE}_{\mathsf{CV}} = \frac{1}{N_{\mathsf{CV}}} \sum_{j=1}^{N_{\mathsf{CV}}} \mathsf{MAE}_j \tag{2.19}$$

where MAE_j is the mean absolute error in cross validation round j, and N_{CV} is the the total number of cross validation rounds. MAE_{CV} was estimated for annual rainfall, daily rainfall

variance and daily rainfall occurrence probability. To simulate the effect of gauge network density on the performances of the three interpolation procedures, I varied the size of the training set, keeping the size of the validation set constant.

(f) Application to the bias correction of time series

I finally illustrate the application of adjusted stochastic model parameters to correct time series through quantile mapping. The method was applied on the TRMM time series recorded above Darchula (1685 m.a.s.l) a rain gauge location in the hilly region of western Nepal (Figure 2.1 (a)). Although the gauge itself features an observation period that overlaps the TRMM time series, records from surrounding gauges were discontinued before the launch of the TRMM satellite, which illustrates the ability of the proposed method to use non-overlapping observations for bias correction. I consider the time series of daily rainfall in September 2005, a period overlapping both rainfall seasons – on average, monsoon ends on September 7th at that location. Similar to the cross validation analysis, stochastic model parameters are adjusted based on information from the neighboring gauges (i.e. excluding Darchula – the verification gauge). TRMM time series are corrected using the adjusted stochastic model parameters as described in Section 2.2.5. The ability of the corrected time series to reproduce the gauged daily rainfall is then assessed and compared to the performance of raw TRMM time series. Finally, for comparative purposes, I also compute TRMM time series corrected by scaling the monthly mean to match the (inverse distance weighted) mean September rainfall observed at surrounding gauges. The latter procedure is very similar to the bias correction operated by NASA on TRMMv6.

2.4 Results and Discussion

2.4.1 Monte Carlo Robustness Analyses

Results from the Monte Carlo analysis are presented in Figure 2.2, showing the results for the four numerical experiments outlined in Section 2.3.1. The outcome of the four experiments was similar: in all cases, combining the ground and satellite data to estimate "true" rainfall resulted in a product that was more robust to errors in either data source. For example, s Figure 2.2 (a), (b) and (c) show how the MAE in annual rainfall estimates responds to different kinds of error sources that impact uncertainty in the gauge data. Figure 2.2 (a) illustrates the effect of elevation bias in the gauge locations, Figure 2.2 (b) shows the effects of observation error at the gauges and Figure 2.2 (c) shows the effects of local rainfall heterogeneities. In each case, and for any given magnitude of the gauge based errors, the MAE computed from bias-adjusted, regionalized estimates with TRMM is much less (often approximately 30% less) than the MAE based on the gauges alone. Conversely, Figure 2.2 (d) assesses the effects of bias in TRMM measurements, and demonstrates that combining gauge data with TRMM stabilizes the MAE in the bias adjusted data even when TRMM itself is biased. Experiments in which both observation errors in gauges and biases in TRMM

were present lead to similar results: the bias adjustment method increased the robustness of the predicted rainfall with respect to the most extreme uncertainty source.

The increased robustness arises due to the near independence of errors in satellite and ground-based rainfall measurements. Since there is not a systematic correlation in uncertainty between these datasets, their joint use stabilizes the bias adjustment method. The results of the Monte Carlo analysis suggest that the proposed bias adjustment procedure is robust to independent errors in the satellite and gauge based observations. This separation of compensating errors is likely to make this data-fusion approach a generic improvement on single-source estimates.

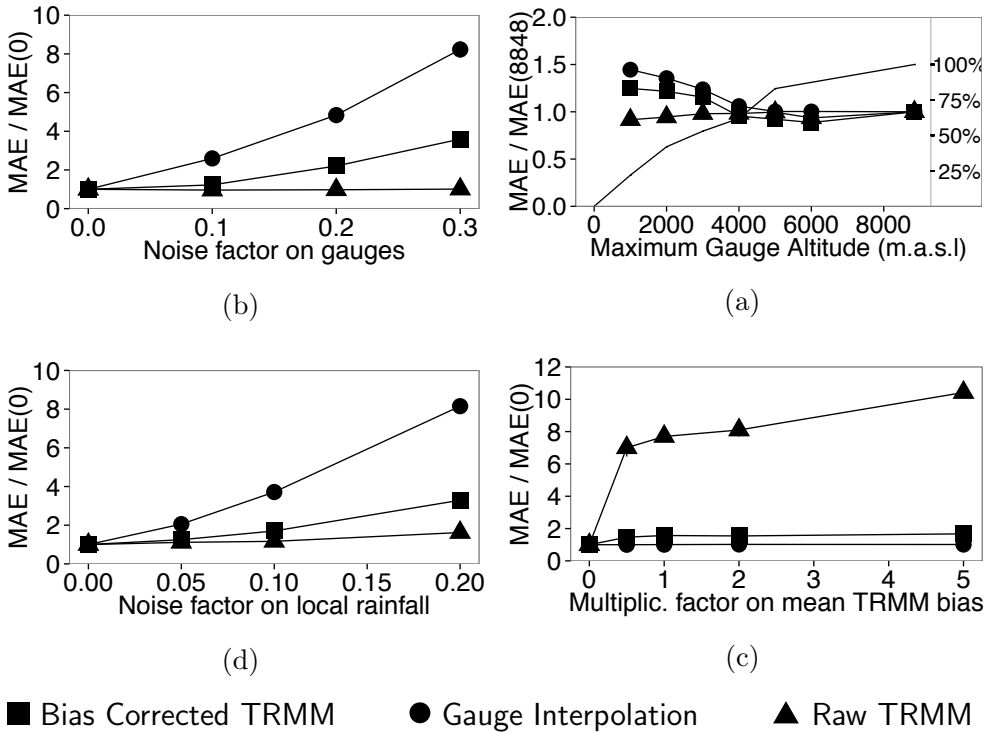


Figure 2.2: Monte-Carlo simulation of the effects of uncertainty sources on the estimated annual rainfall for the bias adjustment method (squares) and the two control methods: unadjusted TRMM (triangles) and interpolation from gauges (circles). The vertical axis represents the mean absolute error on annual rainfall, normalized by its value at the default state described in Table 2.3. (a) Effect of the systematic selection of low altitude gauges: the x axis represents the lower altitude limit set for the randomly selected gauge locations; the graph line without point markers and secondary y axis represent the cumulative altitude distribution of the study area. (b) Effect of the variance of the random observation errors on SMPs observed at synthetic gauges. (c) Effect of the mean amplitude of the TRMM bias. (d) Effect of the variance of local random rainfall variations occurring at a spatial scale smaller than that being captured by the gauge network.

2.4.2 Evaluation of TRMM 3B42 v6 in Nepal

I found large bias in rainfall estimates in Nepal made using TRMMv6. Yearly rainfall was strongly underestimated by the raw TRMMv6 dataset with a mean bias of -539 mm y^{-1} over the study area and a mean absolute error of 580 mm y^{-1}. The 95% confidence interval around the mean bias was 703 mm y^{-1}, suggesting significant spatial variation in the bias, as illustrated in Figure 2.3. TRMMv6 captures large scale rainfall gradients, but misses variations around prominent topographic features. For example, in leeward regions like Mustang TRMM over-estimated the gauged annual rainfall by over 100% (i.e. a relative bias above 1) , while in windward regions like Pokhara TRMM underestimated the gauged annual rainfall by more than 50% (i.e. a relative bias smaller -1). These observations are consistent with previous observations that TRMMv6 fails to reproduce orographic impacts on rainfall [58]. The observed clustering of biases around prominent topographic features leads to spatial heterogeneity in the biases, but also spatial autocorrelation, facilitating the use of kriging techniques for interpolation.

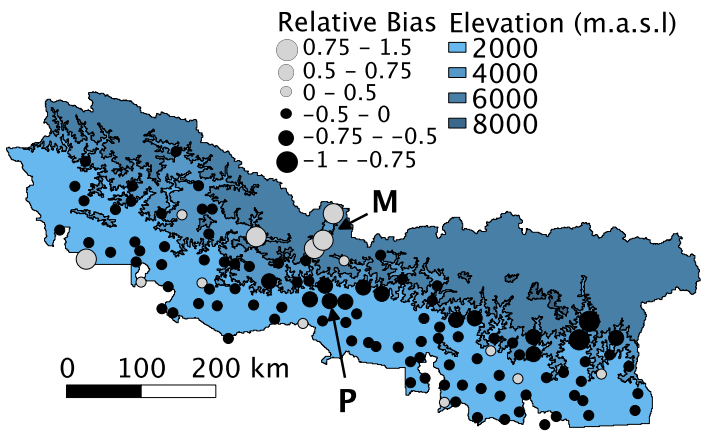


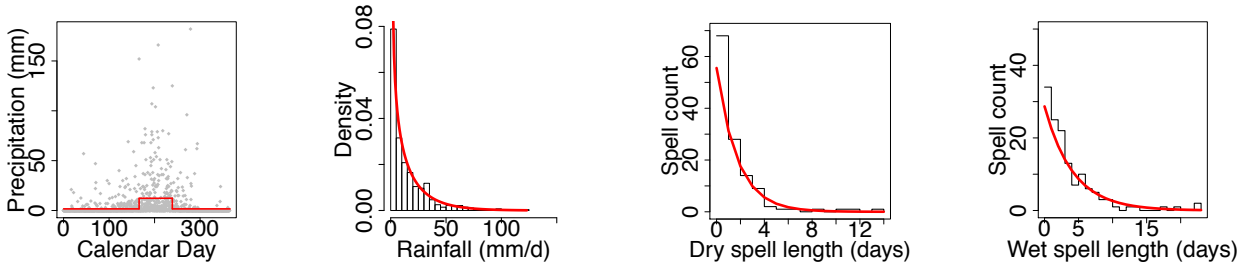
Figure 2.3: Spatial repartition of the TRMM bias on yearly rainfall. The relative bias is calculated by normalizing the observed bias by the yearly rainfall measured at the gauge. A relative bias of -1 means that the average yearly rainfall observed at the gauge is double the value given by the covering TRMM pixel. The large variation and different signs between Pokhara (P) and Mustang (M), two proximate regions separated by the Anapurna Range illustrates the effect of rain shadows on the bias.

The mean, standard deviation and spatial range values for each of the stochastic rainfall characteristics described in Appendix A.1 and calculated from the fitted SMPs are shown in Table 2.4. As shown, TRMMv6 reproduced the duration of the monsoon well: it occurred, on average, between June 7th and September 18th. The beginning and end dates of the monsoon period each had a standard deviation of approximately two weeks across the region.

During the monsoon, 65% of the days were rainy, with average wet spells of 6 days. Only 16% of days were rainy in the dry season, with average dry spells of 12 days. These characteristics were also reproduced by TRMMv6 with a slight overestimation of daily rainfall probabilities. Conditional rainfall intensity was severely underestimated by TRMMv6 which found the intensity to be approximately 50% smaller than that reported by the gauge network. In contrast, Duncan and Biggs [76] found that TRMMv6 under-estimated daily rainfall probabilities and overestimated the rainfall intensity. I attribute the differences between the findings of these studies to a different choice of evaluation metric: rather than evaluating the TRMMv6 product with respect to point gauge data, Duncan and Biggs [76] compared TRMMv6 to interpolated daily precipitation measurements. As discussed in Part (b), errors associated with spatial interpolation of rainfall gauges exceed the error sources in TRMMv6 in regions with low gauge densities. Because of such embedded interpolation errors, the evaluation of TRMMv6 against gridded precipitation stemming from interpolated gauge data is problematic.

2.4.3 Stochastic modeling of Nepalese rainfall

Applying the stochastic model described in Section 2.2.1 to rain gauge data in Nepal lead to a mean absolute error in the annual rainfall of 7.8 mm y^{-1} compared to the observed time series – 0.4% of the region's average annual rainfall of 1754 mm y^{-1}. Evaluating the stochastic model for each TRMM pixel as illustrated for one gauge in Figure 2.4 lead to a mean absolute error of the same order. These results suggest that despite the complexity of Himalayan precipitation processes the local daily rainfall was well described by a simple seasonal parametric model.



(a) (b) (c) (d)

Figure 2.4: Stochastic rainfall parametrization at a gauge in Western Nepal (Lat:29°28', Long:80°32', z=1266m). (a) A step function is fitted to the time series of daily rainfall to determine seasonality. Monsoon starts and ends at calendar days, when the step function is vertical. (b) A two-parameter gamma distribution is fitted on daily rainfall intensity for each season. The fit on Monsoon rainfall is represented in the figure. (c) The distribution of dry spells (here during the dry season) matches a geometric distribution with probability $P_{01}^{(d)}$. (d) The distribution of wet spells (here during the Monsoon) matches a geometric distribution with probability $P_{11}^{(w)}$.

2.4.4 Performance of the Bias Adjustment Method in Nepal

(a) Performance at Gauged Pixels

Adjusting the SMPs at TRMMv6 pixels that contain gauges (Part (d)) reduced the mean error in annual rainfall to -9 mm y^{-1} (90% CI: 30 mm y^{-1}), effectively eliminating it. The mean *absolute* error between gauges and corrected TRMMv6 pixels was reduced by a factor of 45%, from 580 mm y^{-1} to 319 mm y^{-1}. The fact that so much error remains in the MAE indicates significant outlier compensation effects. That is, the biases are eliminated on average, but remain locally important.

(b) Annual Rainfall at Ungauged Pixels

Figure 2.5 shows the results of the cross validation procedure described in Part (e), which illustrates the ability of the bias adjustment method to reproduce yearly rainfall at ungauged locations. Comparing raw TRMMv6 and TRMMv7 to gauges results in MAEs of 580 mm y^{-1} and 404 mm y^{-1} respectively. These values compare to a MAE of 443 mm y^{-1} obtained when interpolating SMPs from all available gauges. Thus, interpolating the existing gauge network in Nepal outperforms TRMMv6 in the estimation of local annual rainfall, but is surpassed by TRMMv7. The MAE related to gauge interpolation increases steadily with decreasing gauge network density, and exceeds that of the unadjusted TRMMv6 for densities below 2 gauges per 10,000 km^2; that is, an average distance between gauges of about 70km. Using all the gauges in the training set (i.e. 80% of the total number of gauges) to adjust the bias on TRMMv6 reduced the mean absolute error in annual rainfall to 391 mm y^{-1}. This represents 22% of the region's average gauged rainfall of 1753 mm y^{-1} estimated through Thiessen polygons (section Part (a)). When considering the perhaps more accurate measure of average rainfall of 1233 mm y^{-1} obtained by adjusting TRMMv6 over the whole study area, the relative error increases to 31%. This includes the effect of errors related to aggregation and spatial interpolation to ungauged TRMMv6 pixels.

(c) Decreasing Returns to Network Density

The error curve for the bias adjustment on annual rainfall is shown in Figure 2.5 (a). This curve flattens and asymptotes to the error curve for the TRMMv7 data when all available gauges are used to correct TRMMv6. This is consistent with the large number of gauges used by TRMMv7 to adjust the remote sensing rainfall estimates. The flattening of the error curve leads to two noteworthy implications. (i) The incremental benefit of adding gauges to the network to adjust TRMMv6 decreases with increasing network density. The curvature appears to be highest at a density of about 2.5 gauges per 10,000 km^2, where the error is decreased to 458 mm y^{-1}, that is 36% of the TRMM-adjusted average rainfall using only 25% of the available gauges. Thus, a relatively sparse network of gauges, integrated in a bias adjustment procedure based on 10 parameters, efficiently corrects TRMMv6 and generates performance levels comparable to TRMMv7. (ii) The hypothetical availability of a dense

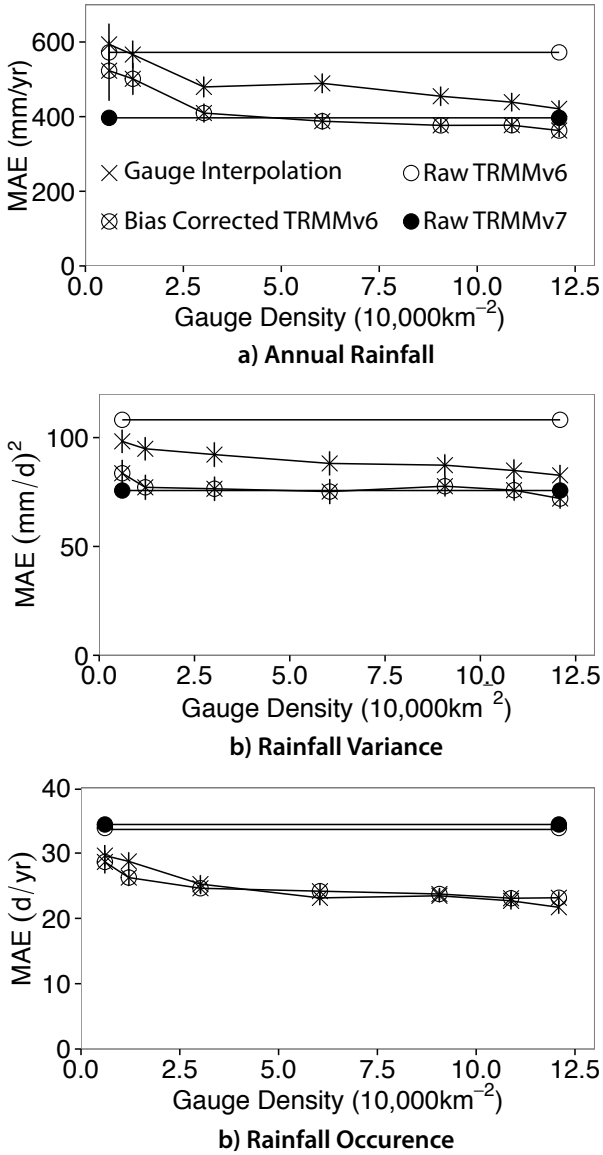


Figure 2.5: Cross validation performances of TRMMv6 (white circles), TRMMv7 (black circles), gauge interpolation (crosses) and bias adjusted TRMMv6 (crossed circles). (a) Mean absolute error on yearly rainfall prediction at ungauged location: bias adjusted TRMMv6 outperforms raw TRMMv6 and gauge interpolation and reaches the performance of TRMMv7 at gauge densities of 6 gauges per 10,000 km^2. (b) Mean absolute error on the variance of daily rainfall: Correcting TRMMv6 leads to equivalent performances than TRMMv7 and both datasets outperform TRMMv6. (c) Mean absolute error on the prediction of the average number of rainy days per year: Gauges outperform both TRMM dataset and improve the performance of bias adjusted TRMMv6.

gauge network – e.g. observed data for every TRMM pixel – to adjust TRMMv6 would result in a non-zero asymptotic error. Indeed, TRMMv7, which is calibrated on 280 gauges, does not outperform a bias adjusted TRMMv6 that uses only 91 gauges. The asymptotic error of 319 mm y^{-1} was estimated using the complete set of available gauges as training and validation sets simultaneously, overriding the aggregation and interpolation steps of the procedure. This residual error is related to omission of local rainfall variations by the coarse resolution of the TRMM satellite and spacing of the Nepalese gauges.

(d) Rainfall Variance and Occurrence Probability

Figure 2.5 (b) and Figure 2.5 (c) show the method's performance at predicting rainfall variance and occurrence using the same cross validation approach as Part (b). For the variance of daily rainfall, the performance of TRMMv7 was reached by correcting TRMMv6 using a small subset of the gauge network. Increasing the density of gauges only slightly improved the performance of gauge-based techniques.

When considering rainfall occurrence, gauge interpolation outperformed both TRMMv6 and TRMMv7 by nearly 30%, with an average error of 21 rainy days per year when all gauges were used. This is consistent with the fact that the TRMM algorithm calibrates remote sensing data using observed monthly mean precipitations, which corrects for average rainfall intensity but fails to adjust biases on rainfall occurrence. The error curve corresponding to the bias adjustment procedure follows the curve related to gauge interpolation, showing that the proposed bias adjustment method successfully corrects rainfall occurrence. Similar to yearly rainfall, the error curve on rainfall occurrence flattens, again suggesting that the incremental benefit of adding gauges to the network to adjust TRMMv6 decreases with increasing network density.

(e) TRMMv7 vs. Bias-Adjusted TRMMv6

Despite the availability in Nepal of high quality TRMMv7 data that successfully represents annual rainfall, the proposed approach to correct TRMMv6 finds its usefulness in its parsimony and its ability to correct hydrologically relevant rainfall statistics using a much sparser gauge network. My approach reached the performance of TRMMv7 in the prediction of annual rainfall using a small subset (90 gauges) of the 280 gauges used in the GPCC dataset to calibrate TRMMv7. Including a stochastic model in the approach allows the daily rainfall to be corrected by adjusting 10 stationary parameters, instead of the 144 monthly means calibrated by the TRMM algorithm for each pixel over a period of 12 years. The proposed method reaches the prediction of rainfall variance and significantly improves that of rainfall occurrence in ungauged locations relative to TRMM v7, using only a subset of the gauges. Finally, I have shown that my method enables even a sparse ground gauge network to correct satellite observations to the same level of accuracy as achieved by monthly-interpolation from a dense network, suggesting that my approach will have applicability in sparsely monitored locations.

(f) Bias correction of time series

Figure 2.6 illustrates the use of bias-adjusted stochastic model parameters to correct TRMM time series through quantile mapping for September 2005 at Darchula (1685 m.a.s.l) in Western Nepal. It is immediately clear from the figure that daily rainfall corrected through quantile mapping (circles) reproduces well the observed time series. With a mean absolute error of 8.3mm over the considered period, the quantile mapping time series outperforms raw (dashed) and rescaled (dotted) TRMM with respective mean absolute errors of 9.5mm and 14.9mm – though the error of rescaled TRMM is likely dominated by gross overestimations of storms on September 15th and 24th.

However, two fundamental limitations of the method are also visible on the figure. (i) Satellites have a limited ability to detect small scale rainfall features, such as the fact that the magnitude of the September 16th storm was lower at the gauge than the pixel average. This limitation is nonetheless common to most remote sensing rainfall estimations and not specific to the proposed method. In fact, unlike mean rescaling, quantile mapping allows representing decreasing biases with rainfall intensities, which prevented the overestimation of the storm of September 24th. (ii) The proposed method addresses the discontinuity of rainfall distribution around zero by generating adjusted rainfall stochastically on days when TRMM records a dry day (8 days days in September 2005), if TRMM overestimates rainfall frequency (i.e. if $F_{adj}^{-1}(P_{TRMM}) > 0$) like in the considered case. The proposed method therefore randomly introduces occurrence errors on certain days (e.g., September 9th), while correcting them others (e.g. September 26th). However, unlike other bias correction approaches, the stochastic method improves the prediction of rainfall frequency, reducing occurrence prediction errors by about 20% at the considered gauge for time series duration of 30 days (5 to 4 errors), 1 year (78 to 57) and 5 years (506 to 401).

2.5 Conclusion

This study explored the potential for bias correction techniques based on stochastic rainfall representations to provide spatially aggregated rainfall data with value for driving hydrological simulations. I have demonstrated that such methods are robust to multiple sources of error and bias in both satellite and ground-based observations of rainfall, and provide robust results for gauge densities as low as 2.5 per 10,000 km^2. I have illustrated that by separating out sources of rainfall observation bias which have different directionalities in different spatial locations, this methodology not only provides a reproduction of rainfall totals which compares to alternative bias correction approaches, such as that applied by NASA for the TRMMv7 dataset; but actually reproduces important statistical features of the rainfall time series, notably the local rainfall variance and rainfall occurrence probabilities, with greater fidelity than obtained from conventional time series bias adjustments.

While a fundamental limitation lies in the inability of satellites to observe small scale rainfall features (a limitation common to other bias adjustment approaches, as shown by the

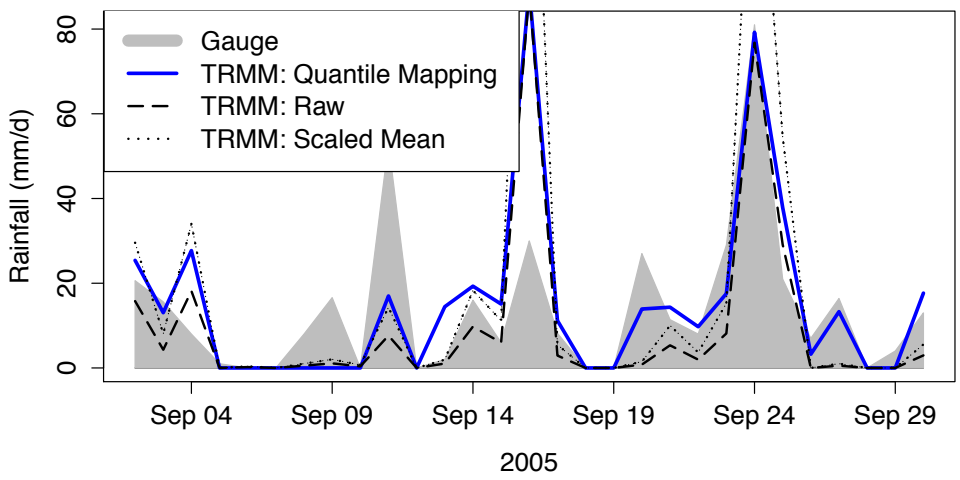


Figure 2.6: Application of the bias-adjusted stochastic model parameters in a quantile mapping procedure to correct daily rainfall time series at Darchula (1683 m.a.s.l) in Western Nepal. The ability of the TRMMv6 time series adjusted with the proposed method (solid) to reproduce gauged values (shaded) exceeded the performance of raw TRMMv6 (dashed) and that of rescaled TRMMv6 (dotted) – i.e. adjusted without stochastic model parameters .

convergence of error estimates between the stochastic approach and the TRMMv7 observations), the proposed method successfully generates parametric distributions of bias-corrected rainfall using a finite number of gauges. Useful application of these results include their use as inputs to frequency domain hydrological models (Chapter 3), the stochastic generation of synthetic rainfall or the correction of remotely sensed time series through quantile mapping.

Thus, the stochastic procedure effectively combines satellite data with sparse rain gauges, providing a robust technique for estimating rainfall properties in minimally-gauged regions, and offering insight into the minimal rainfall gauge network that could be reliably used to understand the spatio-temporal variations in precipitation in mountainous regions.

Chapter 3

Analytical model for flow duration curves in seasonally dry climates

3.1 Introduction

In 2010, about 30% of the world's population lived in areas that experience Mediterranean, Monsoonal or Savanna climates [162], which are characterized by strong seasonality in rainfall. In these climates, annual precipitation is concentrated within one or two wet seasons, followed by extended dry periods. The availability of surface water for ecosystem services (e.g., ecology, domestic supply, irrigation, power generation or sediment transport) during the dry season is strongly dependent on the precipitation volume during the previous wet season and its subsequent discharge from watersheds [123, 191, 13]. For instance in the central Himalayas, up to 80% of the annual rainfall occurs during the three-month long Indian Summer Monsoon (ISM) season. Transient storage of water in fractured bedrock during the rising ISM, and its post-ISM release form the dominant source of dry season streamflow [13], dwarfing the effect of other drivers like evapotranspiration and snow/glacial melt. Because the timing and intensity of precipitation in seasonally dry climates exhibits large inter-annual variation [82] that will likely be amplified by climate change [e.g. 72, 86], both wet season input and dry season water availability are also strongly variable.

In this context, the equitable allocation of seasonally scarce water resources and the design of water-dependent infrastructure are strongly dependent on the reliable prediction of surface water availability and reliability. A key tool used to represent this information is the flow duration curve (FDC): a graphical representation of the probability that a specific magnitude of streamflow is equaled or exceeded [51]. Mathematically, the FDC can be computed as the complement of the cumulative density function (CDF) of daily streamflow. A FDC provides a frequency-domain representation of the daily runoff time series, providing a compact signature of streamflow variability, and its underlying drivers. FDCs are commonly used to estimate water availability for hydropower [e.g. 22], water supply and irrigation [e.g. 61], waste load allocation [e.g. 199], reservoir management [e.g. 6] and environmental health

[e.g. 2]. A comprehensive review of practical FDC applications is presented in Vogel and Fennessey [233].

Despite the utility of FDCs, they are also data-intensive, requiring long-term, on-site daily runoff measurements for their computation [232]. In many parts of the world such data are only sparsely available. The development of techniques to regionalize FDCs to ungauged basins therefore remains an active area of research, especially in arid areas, where the performance of existing methods decreases significantly [51]. Regionalization commonly employs data-intensive statistical approaches, and the calibration of these methods also relies on the availability of long streamflow time series from a large number of representative and well-characterized catchments [e.g. 59, 62]. Thus, traditional regionalization of FDCs remains challenging in truly data-scarce regions. Furthermore, statistical approaches are not able to distinguish the effects of non-stationary climates from those of changing landscape properties, making their application for future flow predictions challenging.

Simple but process-based models of the FDC have the potential to circumvent both of these issues [e.g., 38, 246]. Such models can be developed by extending existing analytical solutions for the probability density function (PDF) of streamflow. In particular Botter et al. [38] analyzed the effect of rainfall forcing on soil moisture and water table recharge. Under the assumption that rainfall occurs as a Poisson Process and that the response time distribution in the water table is exponential, baseflow contribution to streamflow follows a gamma distribution. The resulting streamflow PDF depends on a limited number of physically meaningful parameters related to the stochastic structure of rainfall, and to the soil, vegetation and geomorphologic properties of the catchment. The inherently process-based nature of the approach reduces calibration requirements, allows the effects of changes in climate or the landscape to be independently modeled, and in principle could be driven by remote-sensing observations of rainfall and climate, even where ground-based measurements are sparse, as shown in Chapter 2.

This probabilistic approach has been successfully used to model streamflow PDFs in catchments in the United States, Italy and Switzerland [37, 55, 195, 40]. Two issues, however, need to be addressed before extending the approach to seasonally dry climates. The first issue is general: if these approaches are to be pragmatically used for flow forecasting in multiple basins, then it is likely that the rainfall distribution in at least some of these basins will not be well described by a Poisson process, as seen in Chapter 2. Similarily, deviations from an exponential travel time distribution within the water table are likely in some basins. Although streamflow PDFs for basins with some non-exponential travel time distributions can be determined analytically [39], they are more complex and less analytically tractable than the gamma distribution form. It is therefore valuable to determine how robust the model predictions to deviations from idealized rainfall and catchment properties, and when the simplest PDF description remains valid. The second issue recognizes that previous PDF forecasting with these techniques addressed only seasonal subsets of streamflow time series, where there is an interplay between stochastic water inputs through rainfall, and recessions, through which the excess water is released as streamflow. Yet in seasonally dry climates, where the flow regime during an entire season can be driven by the release of water stored

in the catchment prior to the considered season (e.g.,[13] in Nepal) , the simple analytical model fails because the system does not experience a stochastic Poisson forcing.

In this paper, I extend Botter et al. [38] to link wet and dry season flow generation mechanism and predict annual streamflow distributions in seasonally dry climates. The research is aimed at investigating the following two hypotheses:

H1 Within identifiable limits, the streamflow distribution model described in Botter et al. [38] is robust to deviations from Poissonian precipitation inputs and exponential travel time distributions in the water table.

H2 Streamflow probability distributions during the dry season can be constructed from a deterministic recession relationship with a stochastic initial condition that captures inter-annual variability in wet season characteristics.

I derive analytical expressions for FDCs for seasonally dry watersheds from the superposition of wet season (Hypothesis 1) and dry season (Hypothesis 2) distributions, and evaluate these hypotheses in three different locations with markedly different geologic contexts and distinct climatologies (Figure 3.1): (a) Nepalese Himalayas (topographically complex, deep and shallow soils, Monsoonal), (b) Coastal California (topographically complex, shallow soils, Mediterranean) and (c) Western Australia (flat, deep soils, Mediterranean). I illustrate the practical relevance of the derived analytical method by forecasting the electricity production of two run-of-river hydropower plants in Nepal.

3.2 Methods

3.2.1 Theory

This section describes the derivation of an analytical expression for FDCs in seasonally dry climates. Unless otherwise specified, upper case characters denote random variables and corresponding lower case characters denote realizations of these random variables. The PDF and CDF of a random variable X taken at x are respectively written $p_X(x)$ and $P_X(x)$. Complete, upper incomplete and lower incomplete gamma functions are noted $\Gamma(\cdot)$, $\Gamma_U(\cdot,\cdot)$ and $\Gamma_L(\cdot,\cdot)$ respectively.

The derivation is based on the following key assumptions. Several of these assumptions are necessary for the wet season flow model and reflect those made by Botter et al. [38] in its original derivation. The remaining assumptions pertain to the current extension to seasonally dry systems:

i The watersheds are small enough to that rainfall, soil and vegetation properties can be treated as being spatially homogenous [38].

ii The contribution of glacier or snowmelt is negligible, so that rainfall is the dominant stochastic driver of streamflow [38]. This assumption allows the effects of stochasticity in temperature and solar irradiation to be neglected as drivers of flow variability.

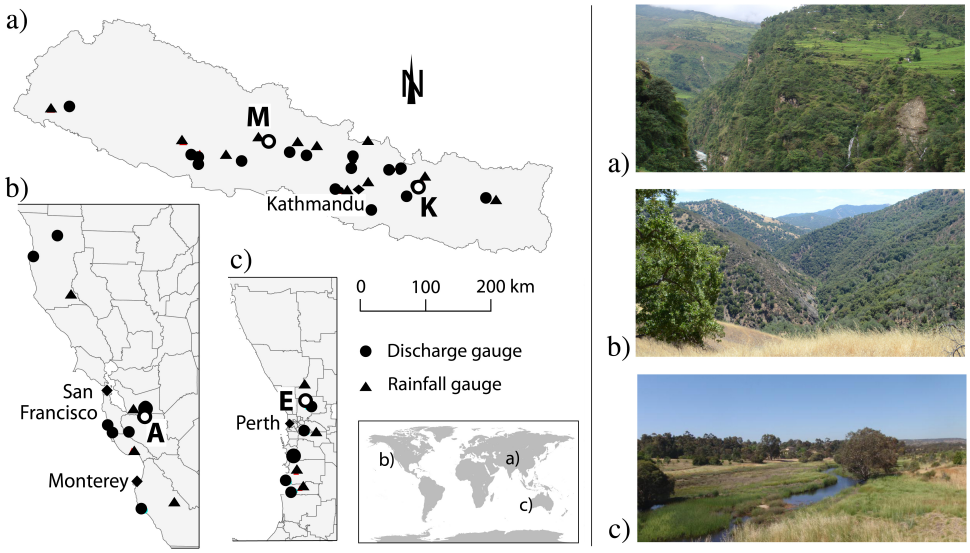


Figure 3.1: Respective locations of the 17, 8 and 6 Discharge gauges and corresponding rainfall gauges in Nepal (a), California (b) and Western Australia (c). Parameters have been estimated using rainfall on the Modi Kohla (M) and Khimti Kohla (K) catchments in Nepal and Ellenbrook (E) in Western Australia because of their small surface area and proximity to a rain gauge. Photo credits: (a) Marc Müller, (b) Gopal Penny, (c) Artemis Kitsios

While the model does not apply to watersheds with where glacial or snowmelt discharge is dominant, it can be adapted for snow-fed basins without seasonal snowpack accumulation [195].

iii There are two distinct precipitation seasons (wet and dry) characterized by a statistically significant difference in mean seasonal streamflow. The duration of the seasons is assumed to be near-constant from year to year, so that the effects of inter-annual variations in the timing of the seasons has minimal impact on the stochasticity of flow.

iv During the rainy season, rainfall is assumed to follow a marked Poisson process with exponentially distributed rainfall intensities. When infiltrated rainfall causes soil moisture to exceed the field capacity of the vadose zone, excess water recharges the water table and is available to generate runoff [38].

v The response time distribution during the wet season is treated as exponential [38].

vi During the dry season, the (minimal) rainfall is assumed to be completely consumed by evapotranspiration in the vadose zone. Thus, this simple model assumes that there is no water transfer between the vadose zone and the water table in the dry season, and only water stored during the previous rainy season drains and feeds the baseflow of the stream.

I assess the sensitivity of the model to violations in three key dynamic assumptions - binary seasonality, the Poissonian character of rainfall in the wet season, and the constant length of each season - through numerical simulations and via case study on watersheds in Nepal, Coastal California and Western Australia.

(a) Steady state rainy-season streamflow distribution

Under these assumptions, the vadose zone censors the frequency of rainfall. If the incoming rainfall is a marked Poisson process, wet season runoff increments also follow a marked Poisson process [38]. The censored runoff frequency λ is related to rainfall frequency λ_P by:[118, 38].

$$\lambda = \eta \frac{exp(-\gamma)\gamma^{\frac{\lambda_P}{\eta}}}{\Gamma_L\left(\lambda_P/\eta, \gamma\right),} \tag{3.1}$$

where $\eta = ET/(nZ_r(s_1 - s_w))$ is the maximum evapotranspiration rate normalized by the root zone storage capacity, and $\gamma = \gamma_P n Z_r(s_1 - s_w)$ is the ratio between the soil storage capacity and the mean rainfall depth $1/\gamma_P$. Z_r, n, s_1 and s_w are parameters respectively representing the depth, porosity, field capacity and wilting point of the root zone and are defined by vegetation and soil type.

Assuming an exponential distribution of travel time in the water table, the steady state distribution of rainy season streamflow Q_w follows a gamma distribution [38]:

$$p_{Q_w}(q_w) = \frac{\gamma_Q^m}{\Gamma(m)} q_w^{(m-1)} exp\left(-\gamma_Q q_w\right), \tag{3.2}$$

$$P_{Q_w}(q_w) = \frac{\Gamma_L(m, \gamma_Q q_w)}{\Gamma(m)} \tag{3.3}$$

with $Q \in \mathbb{R}_{\geq 0}$. The parameter $m = \lambda/k$ describes the ratio between the mean response time $(1/k)$ and mean inter-arrival time $(1/\lambda)$ of recharge events to the aquifer. The inverse of the mean recharge volume is given by

$$\gamma_Q = \frac{\gamma_P}{Ak}, \tag{3.4}$$

with γ_P the mean inverse rainfall intensity and A the contributing area of the watershed. Part (a) assumes small (i.e. smaller than rainfall spatial correlation lengthscale) and homogenous basins, with flow and rainfall measured on time scales larger than the characteristic duration of single rainfall events (e.g., daily streamflow). It does not account for lateral subsurface flow in the vadose zone, losses to deep percolation or overland flow.

(b) Peak flow distribution of the last rainy season storm

Typically, the timescale of the exponential correlation function of the Poisson rainfall process is short enough that streamflow rapidly reaches a steady state following the onset of the wet

season. This steady state streamflow can thus be modeled by a random variable Q_w with a PDF given by Part (a). The dry season recession begins at the realization of Q_w on the last day of the wet season (i.e. at $t = T_w$), which is defined here as occurring at the peak of the last storm of the wet season. The flow Q_0 that represents the initial condition at the beginning of the dry season is thus the sum of two stochastic processes: the flow Q_w discharged before the last wet season storm and the flow increment Δ generated by the precipitation during that storm:

$$Q_0 = Q_w + \Delta$$

With the assumptions above [38], flow increments are independent and exponentially distributed with mean $1/\gamma_Q$.

$$p_\Delta(q_w) = \gamma_Q \mathsf{exp}(-\gamma_Q q_w) \tag{3.5}$$

with $q \in \mathbb{R}_{\geq 0}$. Because the distribution of the sum of two independent random variables is given by the convolution of the distributions of the individual random variables, the PDF and CDF of the streamflow at the end of the wet season can be expressed by convolving p_Δ and p_{Q_w}.

$$\begin{aligned} p_{Q_0}(q_0) &= \int_0^{q_0} p_{Q_w}(q_w) p_\Delta(q_0 - q_w) dq_w \\ &= \frac{\gamma_Q^{1+m}}{\Gamma(m+1)} exp(-\gamma_Q q_0) q_0^m \end{aligned} \tag{3.6}$$

$$P_{Q_0}(q_0) = \frac{\Gamma_L(1+m, \gamma_Q q_0)}{\Gamma(m+1)}, \tag{3.7}$$

The integration in Part (b) is bounded by q_0 because the domain of $p_\Delta(q_w)$ (Equation (3.5)) is non negative. Q_0 follows a gamma distribution with rate γ_Q identical to the rate of Q_w and Δ and with a shape parameter $m+1$. The expectation of Q_0 can be expressed as:

$$\mathsf{E}[Q_0] = \frac{m+1}{\gamma_Q} \tag{3.8}$$

Q_0 is thus an annual stochastic variable representing the inter-annual variability of the intensity of the rainy season (Figure 3.2 (c)).

(c) Modeling recession relations

Following Andermann et al. [13], I hypothesize that dry season streamflow is driven by the release of water stored in the water table during the previous rainy season. I neglect the contribution of snow and glacial melt and of dry season precipitation. During the rainy

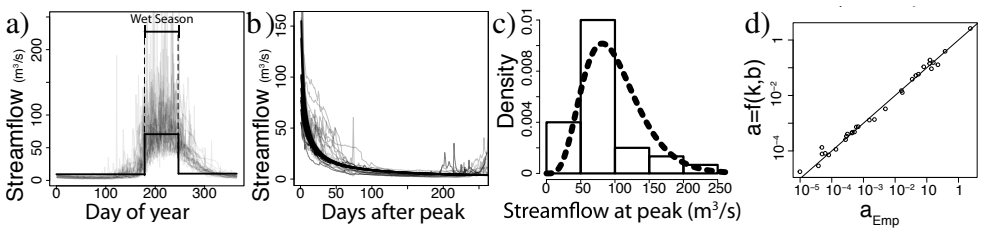


Figure 3.2: a) Estimation of the seasonality parameter at Khimti Kohla. The first vertical segment of the step function represents the median starting day of the wet season flow regime. The length of the following segment represents the median time until the last wet season discharge peak. b) Seasonal recession estimation at Khimti Kohla. Dark lines are fitted recessions for a= 0.00089 and b=2.09 given the observed initial streamflow. c) Empirical histogram of the Khimti Kohla discharge at the end of the wet season for N=30 years. The histogram is overlaid by the analytical PDF computed from Part (b). d) Scatterplot of empirical a parameters estimated on the 31 considered catchments against corresponding values obtained from Equation (3.23). The solid line represents $a_{Emp} = a(k, b)$, $R^2 = 0.98$.

season, frequent recharge events minimize variations of the water table level. Under these circumstances, the Boussinesq equation, which governs water table discharge to the channel, is well approximated by its linearized solution, which is characterized by an exponential travel time distribution [44] and an exponential recession of baseflow in the absence of recharge.

$$Q_w(t) = q_{wo}e^{-kt} \tag{3.9}$$

where parameter k is the linear recession constant and q_{wo} the peak flow at the beginning of the recession.

Once frequent recharge ceases during the dry season, however, the water table undergoes a large transient drawdown, corresponding to non-linear discharge behavior and a power-law response time distribution [44]:

$$\frac{dQ}{dt} = -aQ^b \tag{3.10}$$

Here parameters a and b are assumed stationary and are related to aquifer characteristics (depth, surface area, hydraulic conductivity, porosity and drainage density), and to the boundary and initial conditions for the water table. Integrating Equation (3.10) provides the temporal evolution of the the dry season flow Q_d given an initial discharge q_0[44]

$$Q_d(t) = (q_0^r - art)^{\frac{1}{r}} \tag{3.11}$$

where $r = 1 - b$. Equation (3.11) provides a reasonable description of observed seasonal recessions, as qualitatively shown in Figure 3.2 (b).

(d) Dry season streamflow distribution

Inverting Equation (3.11) allows the time t^* needed for the recession flow to reach the condition $Q_d(t) \leq Q$ to be computed. Then, knowing that $Q_d(t)$ is decreasing, streamflow always meets the condition $Q_d \leq Q$ during the period between t^* and the end of the dry season. This allows the CDF of Q_d conditional on the initial flow Q_0 to be found as (see Appendix B):

$$\begin{aligned} P_{Q_d|Q_0=q_0}(q_d, q_0) &= P\{Q_d \leq q_d \mid Q_0 = q_0\} \\ &= \frac{\mid \{t \in [0, T_d] \mid Q_d(t) \leq q_d\} \mid}{T_d} \\ &= \begin{cases} 1, & \text{if } q_d > q_0 \\ 0, & \text{if } q_d < (K - arT_d)^{\frac{1}{r}} \\ 1 - \frac{1}{T_d}\frac{q_0^r - q_d^r}{ar}, & \text{otherwise} \end{cases} \end{aligned} \tag{3.12}$$

where parameter T_d is the duration of the dry season and where $K = q_0^r$ if $r > 0$ and $K = 0$ if $r < 0$ (implying that $0 < b < 1$).

Knowing the distribution of Q_0 (Part (b)), I obtain the unconditional cumulative density function of dry season flow by applying the law of total probabilities [212]:

$$\begin{aligned} P_{Q_d}(q_d) &= \int_{Q_0} P_{Q_d|Q_0}(q_d, q_0) p_{Q_0}(q_0) dq_0 \\ &= \begin{cases} 1 + \frac{q_d^r \Gamma_1 - \gamma_Q^{-r}\Gamma_2}{arT_d\Gamma(m+1)}, & \text{if } q_d > -(arT_d)^{\frac{1}{r}} \text{ and } r < 0 \\ 1 + \frac{q_d^r \Gamma_1 - \gamma_Q^{-r}\Gamma_2}{arT_d\Gamma(m+1)} + \frac{\gamma_Q^{-r}\Gamma_4 + \left(q_d^r - arT_d\right)\Gamma_3}{arT_d\Gamma(m+1)}, & \text{otherwise} \end{cases} \end{aligned} \tag{3.13}$$

with

$$\begin{aligned} \Gamma_1 &= \Gamma_U(m+1, \gamma_Q q_d) \\ \Gamma_2 &= \Gamma_U(r+m+1, \gamma_Q q_d) \\ \Gamma_3 &= \Gamma_U\left(m+1, \gamma_Q (q_d^r + arT_d)^{\frac{1}{r}}\right) \\ \Gamma_4 &= \Gamma_U\left(r+m+1, \gamma_Q (q_d^r + arT_d)^{\frac{1}{r}}\right) \end{aligned}$$

Full derivations of Part (d) are provided in Appendix B.

(e) Period of Record flow duration curve

Thanks to the memoryless property of the exponentially distributed runoff increments $\Delta^+ Q_{storm}$, the flow Q_0 representing the initial condition at the beginning of the dry season is independent from wet season daily streamflow Q_w. Daily discharge during the wet and daily discharge during the dry season are therefore two independent random variables and the CDF

of streamflow (unconditional on the season) can be expressed as a weighted average of the underlying seasonal CDF (Part (a)):

$$P_Q(q) = \left(1 - \frac{T_d}{365}\right) \cdot P_{Q_w}(q) + \frac{T_d}{365} \cdot P_{Q_d}(q) \tag{3.14}$$

The period-of-record flow duration curve (PoRFDC) is computed by inserting the unconditional CDFs of wet and dry season streamflow (Part (a)) and plotting the streamflow quantiles q against $1 - P_Q(q)$, the complement of the assembled streamflow CDF.

(f) Annual flow duration curves

While the PoRFDC lumps the intra- and inter-annual character of streamflow variations, variability on these two timescales can be separated using annual flow duration curves (AFDC). Empirical ADFCs are constructed for each year using the streamflow records for that year only. The quantile-by-quantile median of the exceedance probabilities from all available AFDCs and their related confidence intervals describe the flow regime of a typical (though hypothetical) year and its inter-annual variation [232]. Both of these can be quite different from the PoRFDC. The information provided by AFDCs is of particular relevance in water resource management applications, where costs and benefits are calculated on a yearly basis, and where the high degree of interannual variability characteristic of seasonally dry climates has direct implication for infrastructure design.

In the proposed model, dry season base flow is driven by an annual stochastic process – the streamflow Q_0 generated following the last wet season storm, and a deterministic intra-annual recession. These features mean that intra-and inter annual streamflow variation can be readily disentangled. During the wet season, I model daily streamflow Q_w as the product of two independent random variables: an annual stochastic index flow $Q_{AF,w}$ and a dimensionless daily streamflow Q'_w. This stochastic index flow approach has been suggested by Castellarin et al. [50] and successfully applied to predict both PoRFDCs and AFDCs in ungauged basins [52] and was recently adapted for intermittent streams [182]. Here, I consider mean runoff as the stochastic index flow for the wet season:

$$Q_{AF,w} = \frac{1}{T_w} \sum_{t=1}^{T_w} Q_{w,t}, \tag{3.15}$$

where $T_w = 365 - T_d$ is the (assumed constant) duration of the wet season. Because all daily realizations $Q_{w,t}$ of wet season base flow follow an identical gamma distribution, the CDF of $Q_{AF,w}$ is a linear transformation of the T_w-fold convolution of the CDF of Q_w given in Part (a):

$$P_{Q_{AF,w}}(q_{AF,w}) = P_{Q_w}^{T_w *}(T_w \cdot q_{AF}) = \frac{\Gamma_L(T_w \cdot m, T_w \cdot \gamma_Q q_{AF,w})}{\Gamma(T_w \cdot m)}, \tag{3.16}$$

where $P_{Q_w}^{T_w *}$ denotes the T_w-fold convolution of the CDF of Q_w. The dimensionless daily streamflow Q'_w during the wet season is obtained by dividing Q_w by its expectation. Its

CDF can be expressed as:

$$P_{Q'_w}(q'_w) = P_{Q_w}(\mu_{Q_w} \cdot q'_w) = \frac{\Gamma_L(m, m \cdot q'_w)}{\Gamma(m)}, \tag{3.17}$$

where $\mu_{Q_w} = \frac{m}{\gamma_Q}$ is the expectation of Q_w.

The CDF of annual quantile n (e.g., $n \in \{0.5, 0.05, 0.95\}$ provides the median AFDC and the bounds of its 90% confidence interval) can be expressed as:

$$P_{Q|n}(q) = \left[1 - \frac{T_d}{365}\right] \cdot P_{Q'_w}\left(\frac{q}{AF^{(n)}}\right) + \frac{T_d}{365} \cdot P_{Q_d|Q_0=q_0}(q, Q_0^{(n)}), \tag{3.18}$$

where $AF^{(n)}$ and $Q_0^{(n)}$ are the n-quantile realizations of AF and Q_0, that is the inverse function (taken at quantile n) of the CDFs of Equation (3.16) and Part (b). $P_{Q'_w}$ and $P_{Q_d|Q_0=q_0}$ are the CDFs representing the intra-annual streamflow variations in the wet (Equation (3.17)) and dry (Part (d)) seasons. Since the function $Q_d(t)$ is monotone in Q_0 (Equation (3.11)), larger realizations of Q_0 lead to larger values of $Q(t)$ everywhere. As a result $Q_0^{(n)}$, the n-quantile realization of Q_0, corresponds to the n-quantile realization of $Q_d(t)$ for all t. Therefore, inserting the resulting $Q_0^{(n)}$ into the conditional dry season CDF (Part (d)) allows analytical expressions for the median AFDC and the considered confidence interval to be derived during the dry season despite Q_0 and Q_d being correlated.

3.2.2 Parameter Estimation

The derived model has six parameters (T_d, λ, γ_Q, a, b, k) related to rainfall and catchment properties. These parameters can be estimated from streamflow or rainfall time series as described below. Summary statistics of the parameters estimated for the case studies are given in Part (a).

(a) Rainfall parameters

The frequency λ and mean intensity $1/\gamma_Q$ of wet season runoff events and the duration T_d of the dry season are all driven by the stochastic structure of rainfall, though λ and γ_Q are also affected by the soil, vegetation and geomorphology of the catchment. These parameters can alternatively be estimated from streamflow or rainfall time series.

Using streamflow, the duration of the rainy season is estimated each year by fitting a step function to the streamflow time series (Figure 3.2 (a)). T_d is then obtained by subtracting the median duration of the rainy season from 365. λ and $1/\gamma_Q$ are estimated by considering the subset S of rainy season days with a positive discharge slope (i.e. day t is selected if $Q_{t-1} < Q_{t+1}$) during the rainy season. I then have

$$\lambda = \frac{N_S}{T_d}, \tag{3.19}$$

$$1/\gamma_Q = \frac{1}{2N_S}\sum_{t\in S} Q_{t+1} - Q_{t-1} \tag{3.20}$$

where N_S is the length of S.

Using rainfall, λ_P can be estimated based on the frequency of rainy season precipitation, and then λ computed via Equation (3.1), drawing on estimated evaporation potential and soil textural properties. The parameter γ_Q can be calculated from the mean intensity of rainfall events (Equation (3.4)), combined with the catchment area and the estimated wet season recession constant, k. T_d can be approximated by fitting a step function to rainfall time series instead of daily streamflow. The resulting dry season duration $T_{d,\mathsf{rain}}$ slightly underestimates T_d (Part (a)) as groundwater recharge causes a time lag between the onset of wet season rainfall and the associated flow response. In the catchments considered in my case study, this lag is correlated at the 99% confidence level to both aquifer storage characteristics (parameter a) and the duration $T_{d,\mathsf{rain}}$ of the dry season. Thus I estimate T_d from rainfall time series empirically by regressing linearly the lags $T_d - T_{d,\mathsf{rain}}$ against $T_{d,\mathsf{rain}}$ and a, which estimation method is provided below.

$$T_d = T_{d,\mathsf{rain}} + h_0 + h_{T_{d,\mathsf{rain}}} T_{d,\mathsf{rain}} + h_a a \tag{3.21}$$

where $h_0 = -129.13$, $h_{T_{d,\mathsf{rain}}} = -0.47$ and $h_a = 146.49$ are the ordinary least squares coefficients of the regression ($R^2 = 0.53$).

3.2.3 Recession parameters

Due to the multiplicity of flow generation processes concurrently represented in a hydrograph, the empirical determination of recession parameters from streamflow time series is a significant challenge and an active field of research (see [e.g. 218, 163, 216]).

Here, I estimate the wet season recession constant k by (i) identifying all recessions (consecutive days of decreasing streamflow) longer than 4 days during the rainy season, and (ii) estimating the parameters of Equation (3.9) by regressing the logarithm of the discharge against time for each recession segment [218]. The recession constant is then obtained by taking the median value of the recorded slope coefficients of the regression.

Dry season recession constants a and b are calibrated stochastically based on Equation (3.11). The initial condition Q_0 is estimated each year as the streamflow value at the last peak before the end of the wet season identified by the fitted step function. The estimates of a and b that minimize the sum of squared errors between the modeled (Equation (3.11)) and observed dry season base flow across all years are determined numerically through simulated annealing [25]. Due to the low frequency of rainfall and overwhelming dominance of baseflow in the dry season, the estimation of a and b through this method appears to be robust to the

	W. Australia $N = 6$	Nepal $N = 24$	California $N = 8$
Catchments			
Number of years	26 (15, 35)	19 (17, 22)	50 (33, 61)
Area [km^2]	306 (22, 802)	813 (512, 2380)	126 (110, 218)
Altitude range [m]	186 (179, 223)	4380 (3120, 6190)	921 (792, 1150)
Max snow cover [% area]	0 (0, 0)	16 (8, 32)	0 (0, 0)
Glaciated watershed	0% (0)	29% (5)	0% (0)
Intermittent flow	100% (6)	6% (1)	50% (4)
Model Parameters			
a [$(m^3/s)^{1-b}d^{2-b}$]	0.13 (0.12, 0.20)	$3.3 \cdot 10^{-4}$ ($4.8 \cdot 10^{-5}$, $9.5 \cdot 10^{-4}$)	0.036 (0.011, 0.052)
b[$-$]	1.60 (1.44, 1.76)	2.40 (2.11, 2.51)	1.86 (1.81, 1.91)
k [d^{-1}]	0.26 (0.19, 0.30)	0.16 (0.12, 0.19)	0.25 (0.19, 0.26)
λ [d^{-1}]	0.33 (0.32, 0.34)	0.42 (0.40, 0.44)	0.24 (0.23, 0.25)
$1/\gamma_Q[mm]$	1.17 (0.64, 2.16)	28.7 (18.5, 60.2)	5.64 (3.48, 15.09)
T_d [d]	299 (295, 299)	280 (273, 286)	306 (305, 306)
Gauged Rainfall			
Annual Rain [mm]	821(719, 925)	2170 (1630, 3230)	616 (479, 769)
λ_p [d^{-1}]	0.52 (0.48, 0.55)	0.66 (0.62, 0.83)	0.48 (0.47, 0.49)
$1/\gamma_p$ [mm]	10.30 (9.63, 10.56)	22.26 (16.87, 27.49)	10.35 (7.42, 13.04)
AR	0.34 (0.33, 0.36)	0.23 (0.17, 0.40)	0.40 (0.36, 0.44)
GS	0.86 (0.78, 0.96)	1.01(0.86, 1.53)	0.74 (0.72, 0.75)
$T_{d,\mathsf{rain}}$ [d]	262 (255, 269)	279 (266, 286)	286 (281, 292)
CV_{T_w} [$-$]	0.25 (0.19,0.28)	0.20 (0.17,0.23)	0.29 (0.27,0.32)
Model Performance (Estimated based on streamflow input)			
	Nash-Sutcliffe Coefficient of log transformed streamflow quantiles		
Period of Record			
Whole year	0.67* (0.60*, 0.77*)	0.90 (0.84, 0.92)	0.97* (0.91*, 0.98*)
Dry season	0.89 (0.86, 0.94)	0.15 (-0.50, 0.60)	0.65 (0.64, 0.73)
Wet season	0.43*(0.22*, 0.60*)	0.78 (0.61, 0.83)	0.95 *(0.84*, 0.96*)
AFDC			
median	0.65* (0.58*, 0.77*)	0.91 (0.84, 0.94)	0.94*(0.92*, 0.96*)
upper CI90	0.23* (0.18*, 0.62*)	0.73 (0.67, 0.80)	0.60 (0.45, 0.65)
lower CI90	0.77* (0.76*, 0.88*)	0.90 (0.85, 0.94)	-0.42* (-5.15*, 0.40*)

* *Intermittant flow at one or more gauges: non positive flow quantiles are omitted*

Table 3.1: Descriptive statistics of catchments by region. In this table, Q_m (Q_{25}, Q_{75}) represent the lower quartile Q_{25}, the median Q_m, and the upper quartile Q_{75} for continuous variables. N is the number of non–missing values. Numbers after percents indicate the number of catchments. Model parameters are estimated from the observed hydrographs. Nash Sutcliffe coefficients are computed on flow quantiles 1/365 to 364/365.

choice of baseflow separation method – here the Lyne Hollick algorithm [163]. The more direct method of regressing the log-transformed rate of change of discharge against the log-transformed base flow [44, 55, e.g.,] resulted in biased estimates of a and b in my case study because the discharge rate is not detectable on a daily scale on a substantial part of the recession.

Alternatively, because the wet season recession constant k and the dry season recession parameters a and b describe the same watershed, they must be related. For consistency between long and short recession behavior, I require that the the power-law recession (left hand side of Equation (3.22)) be approximated by an exponential recession (right hand side of Equation (3.22)) for sufficiently short recession times t:

$$(q_0^r - art)^{\frac{1}{r}} = q_0 e^{-kt} \tag{3.22}$$

Substituting $q_0 = \mathsf{E}[Q_0]$ (Part (b)), the expected value of flow peaks, and $t = 1/\lambda$, the mean duration of recessions during the wet season into Equation (3.22), I solve for a:

$$a = \frac{\lambda}{-r}\left(e^{\frac{-r}{m}} - 1\right)\left(\frac{m+1}{\gamma_Q}\right)^r \tag{3.23}$$

In the analysis, k is estimated independently from a and b using Equation (3.9) and Equation (3.10). However, I test the ability of Equation (3.23) to reproduce the obtained estimate of a in Section 3.3.1.

3.2.4 Evaluation metric

Following Castellarin et al. [49], I compare analytical and empirical FDCs using the Nash-Sutcliffe Coefficient (NSC) applied to the flow quantiles:

$$NSC = 1 - \frac{\sum_{j=1}^{350}\left(\hat{q}_j - q_j\right)^2}{\sum_{j=1}^{N}\left(q_j - \frac{1}{N}\sum_{j=1}^{350} q_j\right)^2} \tag{3.24}$$

where $\hat{q}_j$ and q_j are the empirical and analytical daily streamflows associated with quantile j.

Castellarin et al. [49] use NSC intervals ($]0.75, 1]$: good, $]0.5, 0.75]$: fair, $]-\infty, 0.5]$: poor) to evaluate FDC regionalization methods at ungauged sites. Although this study does not consider ungauged catchments, I use the above intervals as benchmarks to quantify the performance of the model. In order to mitigate the outlier effect of large floods, I take the logarithm of the flow quantiles before computing the NSC. Note that although representative of the overall modeling performance, the NSCs are not necessarily representative of the model's utility in the context of a specific application, which hinges on its ability to predict the duration of particular flows that are exogenously determined by design constraints (e.g., the design flow Q_d in the case of run of river hydropower). Therefore, I use error duration curves [176] to assess the repartition of the errors across flow quantiles. The curves represent

the median 40% and 80% empirical confidence interval of the relative residuals of streamflow values against their duration.

3.2.5 Numerical analysis

Rainfall in several seasonally dry climates, including Nepal (Chapter 2), does not always follow a Poisson process. Similarly, many watersheds exhibit nonlinear recession behavior, which indicates a non-exponential travel time distribution. I evaluate the robustness of the FDC model to the violation of these two assumptions through a numerical analysis, in which I generate streamflow data by routing non-Poissonian synthetic rainfall through a non-linear water table and compare the resulting FDCs to those obtained using Part (a).

To generate synthetic streamflow, I first generate synthetic rainfall time series in which rainfall occurrence is more or less autocorrelated, and in which rainfall intensities follow a gamma distribution, like in Chapter 2. By forcing the first order autocorrelation parameter (AR) to 0, and the shape parameter of the gamma distribution (GS) to 1, these assumptions can replicate a Poisson Process. The synthetic rainfall forces a vadose zone soil moisture balance model with linear losses to evapotranspiration [38, 174]. The resulting water table recharge R replenishes a nonlinear water table yielding a runoff Q described by [129]

$$\frac{d(ln(Q))}{dt} = aQ^{b-2}\left(\frac{R}{Q} - 1\right), \tag{3.25}$$

where a and b are non-linear recession parameters (Equation (3.10)). This initial value problem was solved numerically with the *lsodes* solver [211].

I test the model's robustness to non-Poissonian rainfall and non-exponentially distributed response times (Hypothesis H1) by (i) generating a 10,000 day-long synthetic streamflow record, (ii) estimating the parameters k, λ and γ_Q from the synthetic time series, (iii) constructing the empirical FDC from synthetic streamflow and (iv) comparing it to the analytical FDC (Part (a)) computed from the estimated parameters. The Poissonian character of rainfall was progressively eroded by altering the first order autocorrelation coefficient on rainfall occurrence (AR) away from 0 within the $[-0.3, 0.7]$ range. The exponential character of the rainfall intensity distribution was progressively eroded by altering the shape parameter (GS) away from 1 within the $[0.02, 2]$ interval. Nonlinearity was investigated by increasing the exponent b of the recession relation from 1 (i.e. the linear case of exponential distributed travel times) to a maximum value of 3. The effects of both non-Poissonian rainfall and non-linear water table recessions on the streamflow FDC are expected to decrease with increasing rainfall frequency, and λ_P was therefore varied in the range $[0.2, 0.8]$.

I also assessed the robustness of the model to random inter annual variations in T_w, the duration of the wet season, by simulating 30 years of streamflow over 1000 Monte Carlo runs. At each run, 30 instances of T_w are drawn from a gamma distribution with a given mean (μ_{T_w}) and coefficient of variation (CV_{T_w}). Wet season streamflow time series of length T_w are then generated for each year as described above, and a non linear recession of length $365-T_w$ is finally appended to each year's simulated wet season. Modeling performance is evaluated

by computing the NSC of the modeled PoRFDC, median AFDC and 5th percentile AFDC (which all assume a constant T_w) against corresponding simulated streamflow distributions. I investigate the effects of fluctuations in the mean and variability of T_w by varying μ_{T_w} in the $[40, 120]$ interval and CV_{T_w} in the $[0, 4]$ interval – the further CV_{T_w} is from 0, the larger the random inter-annual variations in the duration of the wet season.

3.2.6 Case Studies

I rely on observed streamflow data to evaluate whether dry season streamflow PDF can be constructed from a deterministic recession relationship with a stochastic initial condition (Hypothesis H2). I used daily streamflow observations from from 24 catchments in Nepal (Nep) [105, 66], 8 in Coastal California (CA) [224] and 6 in Western Australia (WA) [235] with between 11 and 76 (mean: 18) years of daily streamflow records. The location of the gauges is shown in Figure 3.1, and Part (a) provides a summary of the relevant catchment characteristics and rainfall statistics from daily rainfall time series recorded by precipitation gauges [105, 66, 47, 46] closest to the catchment centroids. Nepalese watersheds are subject to the seasonality of the Indian Summer Monsoon and to the complex topography and variable soil depths of the Central Himalayas. Californian and Southwestern Australian watersheds are subject to a highly seasonal Mediterranean climate with significant winter precipitation. While Californian watersheds are characterized by shallow soils and complex topography, Australian catchment are flat with deep soils.

The modeling approach is tested by comparing empirical PoRFDCs and the median and 90% empirical confidence interval on AFDCs generated from dry season streamflow to their analytical counterparts. Because the main stochastic driver of dry season streamflow (Q_0) is an annual random process, the evaluation of AFDCs is done on the gauges with more than 20 years of available data. Empirical PoRFDCs (using both wet and dry season streamflows) are finally compared to their analytical counterparts.

The predictive ability of the method is assessed by reproducing the above analyses using rainfall (instead of streamflow) data to estimate λ and γ_Q. In order to limit the effect of spatial rainfall heterogeneities, which can be significant in the Himalayas , as shown in Chapter 2, the rainfall-based parameters are computed for a subset of three catchments (two in Nepal and one in Western Australia) that are small and where the catchment centroid is close ($< 20km$) to a rainfall gauge (Section 3.2.6).

3.2.7 Application: Estimation of electricity production using flow duration curves

One final analysis was conducted to illustrate the potential value of the modeling approach for infrastructure design, including an analysis of error not only in the FDC itself, but in the propagation of any such errors into infrastructure design criteria. Flow durations have a direct impact on energy production from run-of-river hydropower facilities. The energy produced by a hydropower plant in a period T is the time-integral of instantaneous power

	Khimti Kohla (K)	Modi Kohla (M)	Ellenbrook (E)
Catchments			
Location	Nepal	Nepal	W. Australia
N_{yr}	30	21	38
Area $[km^2]$	310	142	581(90)
Z_{Rge} $[km]$	3.8	4.5	0.2
Snow [%]	12.1	6.5	0.0
AI $[-]$	1.92	2.87	0.34
Interm. $[d/yr]$	0	0	142
P_{yr} $[mm/yr]$	2230	3350	653
λ_P $[d^{-1}]$	0.90	0.91	0.49
γ_P^{-1} $[mm]$	18.8	28.9	8.3
AR $[-]$	0.23	0.51	0.33
GS $[-]$	1.09	1.23	0.76
$T_{d,\mathsf{rain}}$ $[d]$	277	286	252
CV_{T_w} $[-]$	0.19	0.19	0.24
Model Parameters			
a $[(m^3/s)^{1-b}d^{2-b}]$	0.00089	0.0015	0.129
b $[-]$	2.09	2.14	1.78
k $[d^{-1}]$	0.11	0.18	0.30
λ $[d^{-1}]$	0.44	0.44	0.34
γ_Q^{-1} $[mm]$	18.9	14.1	2.3
T_d $[d]$	276	285	300
Model Performance (Estimated based on precipitation input)			
	Nash-Sutcliffe Coefficient of log-transformed streamflow quantiles		
Period of Record			
Whole year	0.97	0.95	0.69*
Dry season	0.92	0.85	0.22*
Dry seas. baseflow	0.97	0.93	0.77*
Wet season	0.83	0.93	0.82
AFDC			
Median	0.98	0.96	0.70*
CI90 (top)	0.75	0.76	0.05*
CI90 (bottom)	0.96	0.99	0.72*

* *Intermittant flow: non positive flow quantiles are omitted*

Table 3.2: Rainfall-Estimated catchments. In this table, N_{yr} is the number of complete years with available observations, Z_{Rge} the altitude range of the catchment, D_{RnGge} the distance between the catchment centroid and the nearest rain gauge, $Snow$ the maximum snow covered area ratio, AI is the aridity index P/PET, $Interm$ the average number of days per year without flow and P_{yr} is the mean yearly rainfall, λ_P is the rainfall frequency during the wet season and γ_P the inverse of mean rainfall intensity. AR is the first order auto-correlation coefficient of rainfall occurrence, GS is the shape parameter of a gamma distribution fitted on rainfall intensity, $T_{d,\mathsf{rain}}$ the median dry season duration estimated from precipitation time series and $\mathsf{SD}(T_{d,\mathsf{rain}})$ its standard deviation. Model parameters are estimated based on gauged rainfall, assuming actual evapotranspiration values of $2.1mmd^{-1}$ (Nepal) and 1.6 mmd^{-1} (W. Australia), and soil moisture capacities of 16mm (Nepal) and 150 mm (W. Australia). The approximate catchment area effectively contributing to the Ellenbrook streamflow is given in parenthesis. Nash Sutcliffe coefficients are computed on flow quantiles 1/365 to 365/365. The model performance reported for Ellenbrook is based on the $90km^2$ of catchment contributing to streamflow, not the $581km^2$ topographic watershed

generated from the available discharge:

$$E = \rho g H \int_0^T \eta(Q^*) Q^*(D) dD \tag{3.26}$$

where ρ is the density of water and g the acceleration due to gravity. H is the (assumed constant) net hydraulic head drop across the turbine. $Q^*(D)$ is the average discharge passing through the turbines on day D and is related to the FDC $Q(D)$, the design capacity Q_{DC} and the minimal flow discharge prescription MDF [22].

$$Q^*(D) = \begin{cases} Q(D) - MDF, & \text{if } Q(D) - MFD < Q_{DC} \\ Q_{DC}, & \text{otherwise} \end{cases} \tag{3.27}$$

The streamflow effectively used to generate electricity is thus bounded by the design capacity of the turbine. Predicting electricity production therefore requires correctly estimating by the duration of the lower quantiles (i.e. $Q(D) < Q_{DC}$) of the FDC. Turbine efficiency $\eta(Q^*)$ may be modeled as a step function with constant value of η_T above a discharge threshold $\alpha_T Q_d$, below which no electricity is produced. If N such turbines are combined, the plant will have an overall design flow of $N \cdot Q_{DC}$ and will function with an efficiency of η_T for flows above $\alpha_T \cdot Q_{DC}$.

Section 3.3.3 displays the design parameters of two hydropower plants located directly downstream of the Nepalese streamflow gauges included in the analysis in Section 3.2.6. The practical relevance of the FDC model derived here is tested by propagating the modeling errors from FDCs to electricity production estimates. Firstly the long term annual electricity production is evaluated based on the period of record analytical FDC determined from rainfall parameters, and compared to production estimates based on empirical FDC. The (rainfall estimated) median AFDC and the lower bound of the 90% CI are then used to estimate the electricity production in a typical and particularly dry year.

3.3 Results and Discussion

3.3.1 H1: Numerical Analysis

Results from the Monte Carlo analysis are presented in Figure 3.3, showing the outcomes of the three numerical experiments described in Section 3.2.5.

With Nash Sutcliffe coefficients (NSC) above 0.97 within the range of parameters encountered in my case studies (i.e. $\lambda_P > 0.4$, $GS > 0.7$, $AR \in [0, 0.5]$), the first experiment showed that non-Poissonian rainfall has a negligible effect on the model's ability to reproduce wet season streamflow (Figure 3.3 (a) and (b)). With effects on the NSC below 0.1, auto-correlated rainfall occurrence has little effect on model performance, especially for high rainfall frequencies. Gamma distributed (rather than exponentially distributed) rainfall intensities significantly affect the model performance for shape factors < 0.5, as NSC tends towards $-\infty$ when GS tends toward 0.

In line with Ceola et al. [55], the second experiment (Figure 3.3 (c) and (c)) showed that non-exponentially distributed water table travel times have a significant effect on the model performance. The effect was especially visible when the power coefficient b was above 2, denoting a hyperbolic storage-discharge relation [129]. However, as hypothesized in Part (c), the effect of non-linearities decreased with increasing rainfall frequency. Although most catchments are characterized by $b > 2$ (Part (a)), the high frequency of rainfall allowed modeling the catchments as linear reservoirs during the rainy season, resulting in a good match ($R^2 = 0.92$) between the non-linear recession parameters a obtained from Equation (3.23) and the corresponding empirically estimated values of a (Figure 3.2 (d)).

The third experiment showed that within the range of seasonality parameters encountered in my case studies (i.e. $CV_{T_w} < 0.5$ and $\mu_{T_w} \in [60, 100]$), stochastic variations in the duration of wet seasons do not have a significant impact on the performance of the any of the FDC models. The lower-quantile-AFDC (in grey on Figure 3.3 (f)) appears more sensitive to random variations in T_w which affect modeling performance for CV_{T_w} values as low as 0.5 at a mean T_w of 40 days. In all the other considered cases (median AFDC, PoRFDC), variations in T_w only seem to have a significant effect on the modeling accuracy for CV_{T_W} values above 1.

3.3.2 H2: Case studies

(a) Hydrograph-based analysis

Nash-Sutcliffe coefficients for FDCs from the 38 catchments are presented in Part (a), using direct hydrograph observations to estimate the model parameters λ and γ_Q, and thus excluding errors introduced by precipitation observation and the vadose zone model parameterization. The corresponding error-duration curves, which display the repartition of the relative errors across flow durations, are presented in Figure 3.4.

Most period-of-record FDCs are well reproduced by the model with median logNSC above 0.65. With a median logNSC of 0.43, the wet season in Nepal is a notable exception. There the fit is significantly better on high flow (non log transformed NSC=0.79). The observed underestimation of low flows (Figure 3.4 (a)) can be attributed to a strong deviation from the exponential response time assumption in Nepalese basins (median b=2.40). Yet the errors generally do not propagate to non-seasonal FDCs because dry season streamflow is driven by the last wet season peak, which appears to be well estimated, as seen on Figure 3.4 (b) at duration 0. A second exception to the generally good PoRFDC estimates arises for the dry season FDC in WA (median logNSC=0.15). With a median (non log transformed) dry season NSC of 0.70, the poor performance in WA is explained by the intermittent nature of the streams and the exaggerated impact of very low flows on the logNSC.

Predictions of the median dry season AFDC overall were good with median logNSC above 0.64. With a median logNSCs of 0.75, the lower bound of the 90% CI was well reproduced in Nepal and WA, but not CA, where flow quantiles were overestimated (and caused the large spread of error observed in Figure 3.4(e)). The model reproduced the upper bound of the

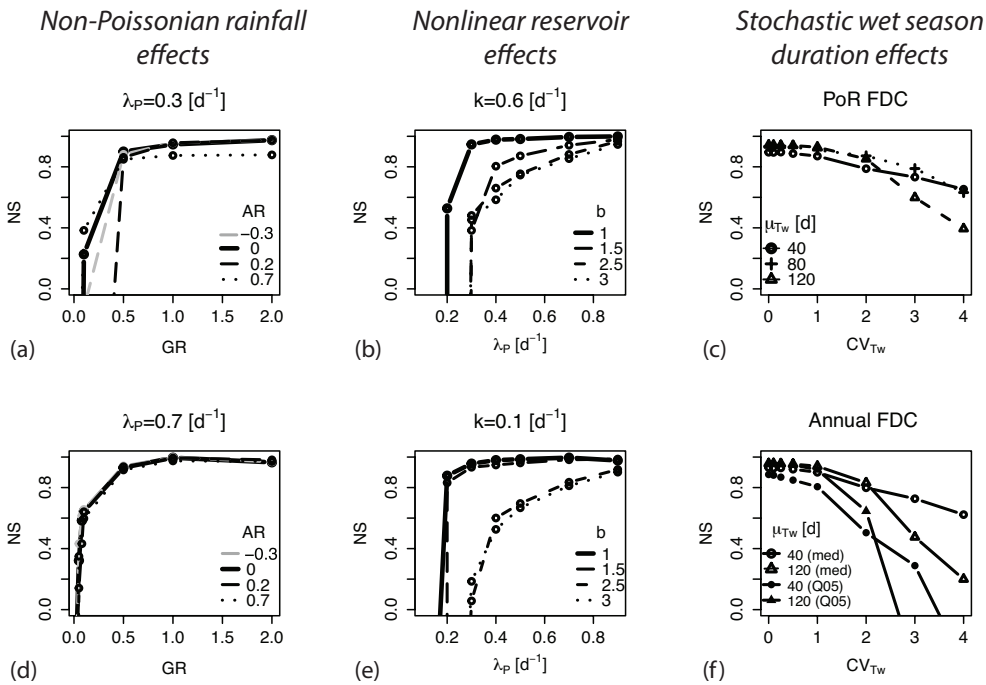


Figure 3.3: Results of Monte Carlo analyses showing the effects of non-Poissonian rainfall (a and d) and non-exponentially distributed water table response times (b and e) on the model performance (as measured by the NSC) for wet season streamflow. Subfigures (c) and (f) show the effect of stochastic wet season durations on the NSC of PoRFDCs (c), of median AFDC (f, white symbols) and 5th percentile AFDC (f, black symbols) . Default parameters: $A = 6500km^2$, $ETP = 3.5mm \cdot d^{-1}$, $nZ_r(s1 - sw) = 180mm$, $\gamma_P = 0.06mm^{-1}$,$\lambda_P = 0.73d^{-1}$, $k = 0.6d^{-1}$, $b = 1$, $AR = 0$, $GS = 1$.

90% CI with median $logNSC > 0.6$ in Nepal and California and a median logNSC dropping to 0.23 in WA.

(b) Rain-based analysis

The FDCs related to Khimti (Nepal) and Ellenbrook (WA), the two catchments selected for the rain-based analysis are presented in Figure 3.5. Unlike the results summarized in Part (a), the parameters T_d, λ and γ_Q of the analytical FDCs were calculated from the seasonality, frequency and intensity of gauged rainfall, taking locally reported values for actual evapotranspiration and available soil moisture capacity (Section 3.2.6). The examples in Figure 3.5 were selected to illustrate model performance for a subset of catchments deviating from standard model assumptions: (1) FDC estimation in a non-homogeneous, arid catchment; (2) Effects of poorly marked seasonality with significant dry season rainfall and (3) Spatially heterogenous rainfall. I explore the consequences of these deviations and some opportunities to adapt the simple model.

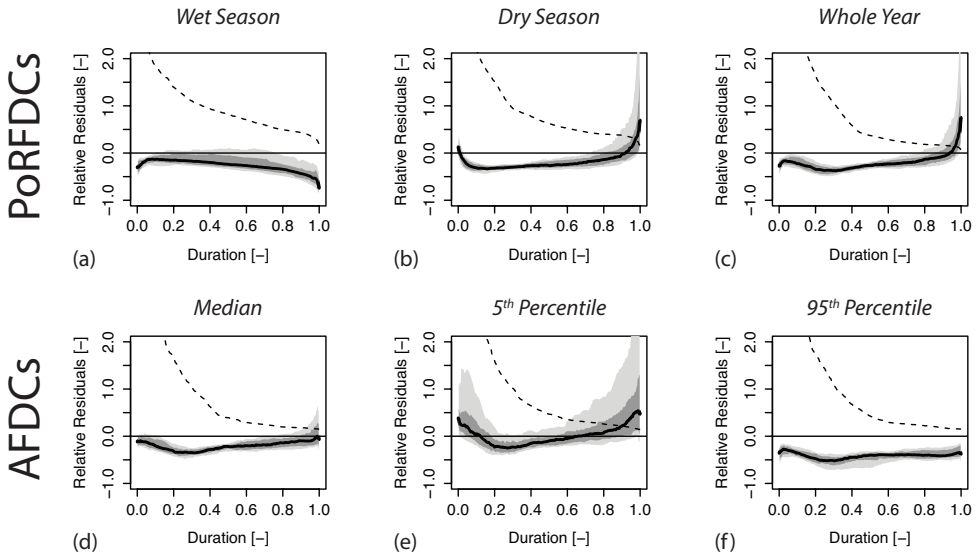


Figure 3.4: Error- duration curves. The thick line represents the median of the relative error per streamflow duration across all sites. The grey nested bands contain 40% and 80% of the relative errors. The hashed line is the median FDC normalized by the mean discharge. PoRFDCs are presented in the first three subfigures (a: wet season, b: dry season and c: whole year) and AFDCs are presented in the last three subfigures (d: median AFDC, e: 5th percentile AFDC and f: 95th percentile AFDC).

Example 1 concerns FDC estimation in arid catchments, as exemplified by Ellenbrook (WA) (Figure 3.5 (e) to (h)). There, the over-estimation of wet season streamflow (Figure 3.5, (e) dotted line) is likely attributable to geological heterogeneities in a catchment where a significant proportion of the catchment area recharges groundwater rather than contributing to surface flow [20]. Reducing the modeled catchment area to the 20% of the catchment thought to supply the majority of baseflow [20] increased the model performance dramatically (solid line).

Example 2 concerns the assumptions that no runoff occurs during the dry season. This leads to an underestimation of the duration of high flows during the dry season in locations poorly marked rainfall seasonality. This effect is particularly visible in Ellenbrook (WA) (Figure 3.5 (f)), where the underestimation of dry season streamflow propagates to the PoRFDC. However, if the model output is compared to base flow (gray diamonds) rather than total flow (grey dots) then the performance metrics drastically improve in both catchments (Figure 3.5 (b and f)). This assumption may also explain the underestimation of the higher bound of the 90%CI of the AFDCs (Figure 3.5 (e and h)): dry season precipitation is likely to occur in particularly wet years. Conversely, dry-season precipitation is less likely to occur in particularly dry years, leading to the improved fits for the lower quantile AFDC.

The final example relates to the mis-estimation of parameters in areas with spatially heterogeneous rainfall, which likely explains the underestimation of wet season flows in

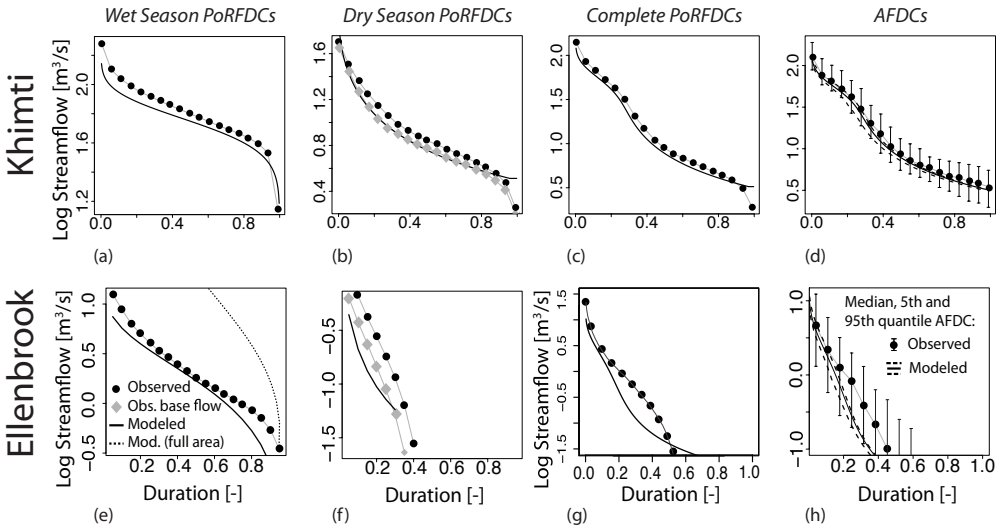


Figure 3.5: Empirical and analytical flow duration curves at Khimti, Nepal, ((a) to (d))and and Ellenbrook, Western Australia ((e) to (h)). PoRFDCs are given for the wet ((a) and (e)) and dry ((b) and (f)) seasons, and for the whole year ((c) and (g)). Median, 5th and 95th quantile AFDCs are displayed in (d)and (h). The modeled wet season curve at Ellenbrook (e) represents the analytical FDC obtained from rainfall over 1/5 of the topographic catchment area - the corresponding FDC over the whole catchment area is given in dotted lines. Grey diamonds in (b) and (f) represent the empirical FDC obtained from the dry season base flow time series that was filtered using the Lyne Hollick algorithm. [163].

Khimti by rainfall-estimated parameters (Figure 3.5 (a)). While the model performs well for streamflow-derived statistics, estimating catchment-averaged rainfall parameters from rain gauge observation is challenging because of the complex topography, as seen in Chapter 2.

Despite these limitations, the analytical FDCs derived from rainfall-estimated parameters reproduced their empirical counterparts well, with most logNSCs above 0.75 (Section 3.2.6). In particular, while neglecting snowmelt contribution, the model performed well for all the FDC types at Khimti ($logNSC > 0.75$) despite significant snow/ice cover (12%). The model was also able to reproduce specific hydrologically significant quantiles like the change in streamflow regime, visible at a duration of 0.3 on the PoRFDC and median AFDC in Khimti (Figure 3.5 (c) and (d)), and the duration of the absence of flow during the dry season in Ellenbrook (Figure 3.5 (f)). Finally, the estimation method (hydrograph or rainfall) had little overall impact on the good performance of the model.

	P $[MW]$	Q_{DC} $[m^3s^{-1}]$	H $[m]$	Turbine Equipment	Observed Prod. $[GWhy^{-1}]$			Modeled Prod. $[GWhy^{-1}]$		
					PoR.	AFDC		PoR	AFDC	
						med.	5th%-ile		med.	5th%-ile
Khimti 1 (K)	60	10.75	684	5 x Pelton: $\eta_T = 0.82$, $\alpha_T = 0.24$	409	409	338	391	393	379
Modi Kohla (M)	13	26	67	2 x Francis: $\eta_T = 0.75$, $\alpha_T = 0.05$	53.8	53.5	45.9	46.0	47.0	41.0

Table 3.3: Design parameters and electricity production estimates of two Nepalese run-of-river hydropower plants. The observed production is computed from observed streamflow records. The modeled production is estimated using the analytical flow duration curve models based on rainfall records. The long term, median and 5th percentile annual productions are calculated based on the PoRFDC, and median and 5th percentile AFDC respectively.The assumed minimum flow discharged (MFD) is 0.5 m^3s^{-1}.

3.3.3 Application: Estimation of electricity production using flow duration curves

Electricity production estimates for the two Nepalese hydropower plants are presented in Section 3.3.3. The long term average, median and 5th quantile of the annual power production are estimated using empirical FDCs (*Observed Prod.*). These empirical results are then compared to the corresponding values obtained from the analytical FDCs (*Modeled Prod.*) based on parameters estimated using rainfall time series. Electricity production is generally underestimated for the catchments (particularly at Modi Kohla). This is related to the underestimation of the duration of high flows (Figure 3.5 (a)) caused by the lower seasonality of the region, which experiences pre- and post-Monsoon storms. Conversely, electricity production is overestimated on dry years (5th percentile) at Khimti Kohla because from the inability of the model to reproduce the low flow conditions of the stream in dry years, when low discharge prevents generation for two weeks. This error source is amplified when considering production variability, that is the difference in production between a typical (i.e. median) and dry (i.e. 5th percentile) year. The error on production variability reaches 80% in Khimti and 30% at Modi. Overall, however, the predictions of annual power production were excellent, with errors below 15% for long term average production and below 12% for annual production quantiles.

3.4 Discussion

3.4.1 H1: The wet season FDC model is robust to deviations from key underlying assumptions

Although the derivation of the original model relies on exponentially distributed response times, rainfall intensities and rainfall interarrival times, my results show that predictions of wet season streamflow PDFs are relatively robust to small deviations from these assumptions. Yet the combination of hyperbolic storage-discharge relationships and low rainfall frequency

reduced the model predictive ability. This situation arises in the strongly nonlinear recessions in Nepal, where the model overestimates the flashiness of wet-season recession. There, this effect was nonetheless mitigated by the high rainfall frequency occurring during the ISM and had little effect on modeling accuracy beyond the wet season.

The assumption of a spatially heterogeneous watershed was violated in Ellenbrook (WA) in which regions with a sandy geology do not generate streamflow. With known geology, these effects could be satisfactorily corrected by adjusting the contributing area of the watershed.

Finally, the challenges associated with estimating catchment-scale effective rainfall statistics is illustrated by the case of Khimti (Nep), in which the model performs well for streamflow-derived statistics, but cannot reproduce these statistics based on the single rain gauge measurement in the topographically complex Himalayan catchment.

Thus, the modeling approach performs well in gauged basins, and holds promise for future application to ungauged basins.

3.4.2 H2: The dry season FDC can be modeled as a deterministic recession relationship with a stochastic initial condition

I modeled dry season streamflow as an annual stochastic process driven by the intensity of the previous wet season and subject to a deterministic recession. Empirical dry season FDCs in my case studies suggest that this simple model captures key flow behavior in seasonally dry regions. The high rainfall seasonality characteristic of these regions is a key prerequisite for the model to be applicable, as evidenced by its poorer performance during the dry season at Ellenbrook (WA). There, a 48 day lag has been observed between the beginning of wet season precipitation and a persistent streamflow rise. Runoff increments during that time lead to the underestimation of high flows during the dry season.

While numerical simulations have shown that the model is robust to fluctuations in the timing and duration of the wet season, unaccounted fluctuations in the frequency and intensity of dry season storms affects modeling performance in watersheds with weaker rainfall seasonality. This, along with the likelihood that during extreme rainfall events flow generation processes bypass the water table, also explain the model's poor prediction of the higher CI bound on the AFDCs.

Generally, qualitative results (Figure 3.2) and the overall good modeling performance on long term and annual FDCs support the utility of the proposed model and point towards water table discharge as the main mechanism for dry season flow production in the considered catchments. Intra-annual flow variations are deterministically driven by the water table, while inter-annual variations are stochastically forced by wet season rainfall. In the Nepalese context, this supports previous findings [13] pointing towards the storage in the fractured bedrock and subsequent release of large volumes of water from the previous monsoon as a key flow generation mechanism. Unlike existing models for seasonally dry climates [e.g., 182, 177], where an atom of probability associated to zero flow is assigned to the entire dry

season, my approach allows accounting for that important mechanism. In Nepal, integrating such a seasonal recession in my model improved the median NSC on period of record FDCs by 0.18, compared to an alternate model where a constant dry season flow of zero was assumed. My results also support the conclusion that the contribution of snow and glacial melt to streamflow variability is small in the Central Himalayas – up to 10% of the streamflow volume according to Chalise et al. [56].

Although the method allowed FDCs to be modeled in relatively dry areas like California, modeling discharge in arid climates remains a significant challenge [51]. Under such conditions, temporal shifts and/or spatial heterogeneities can have a very significant effect on streamflow. Thus the Ellenbrook catchment (WA), where local geological conditions affect the ability of the hillslope to generate streamflow [20] and where a 20% decline in rainfall since 1970 has led to a 65% decline in average streamflow [222], likely represents a limiting case with respect to the applicability of the proposed model in arid catchments. Nonetheless, period-of-record FDCs were successfully modeled at all locations with most median NSC coefficients above 0.75 – the *good fit* benchmark of 0.75 in Castellarin et al. [49].

3.4.3 Practical Relevance

Being able to estimate the inter- and intra-annual variation of streamflow has considerable practical importance, notably to inform water resources and ecosystem management policies [185, 186] and hydropower operations. This was particularly evident in the run-of-river power generation case-study, where electricity production can decrease by up to 20% in dry years, potentially affecting the short term financial sustainability of the infrastructure. My analysis of two Nepalese run-of-river power plants has shown that a significant fraction of the inter annual variation of electricity production can be reproduced using rainfall statistics and recession constants to model the inter-annual variability of wet season streamflow. The model allows the ensuing cash-flow variability to be accounted for during the design phase of the infrastructure, which is currently typically based on period of record FDCs and assumes constant annual revenues [e.g. 107, 192, 22].

A further advantage offered by the process-based nature of the model lies in its ability to disentangle the effects of changes in climate and and landscape on streamflow dynamics. The proposed approach offers an appealing alternative to extend to seasonally dry climates existing models relating catchment storage dynamics to nutrient transport [e.g., 23], landscape characteristics [e.g., riparian width: 155] or ecological dynamics [e.g., plant pathogen risks: 221].

Finally, although not explicitly addressed in this study, the model offers a promising approach to the regionalization of FDCs to ungauged catchments because it relies on a limited number of physically observable parameters. Many of these parameters (e.g., catchment areas, rainfall, evapotranspiration, soil type) are directly and globally available as gridded datasets. However, the study also showed that the model is sensitive to spatial heterogeneities in catchment characteristics and to the accurate computation of catchment-scale rainfall statistics. These effects, in addition to the propagation of errors from gridded datasets,

on the model's performance in ungauged catchments are yet to be assessed. Nonetheless, as indicated by the excellent estimation of run-of-river hydroelectricity production, the modeling approach is apparently well suited to support large-scale site suitability analysis for water infrastructure development [e.g. 248, 130, 135, 137]

3.5 Conclusion

In this study I derived an analytical expression for the FDC of streams in seasonally dry climates. The approach can be successfully applied in a wide range of conditions that are observed in seasonally dry climates and is relatively robust to deviations from the assumptions utilized in the development of the model theory. The process-based nature of the proposed model offers numerous advantages, including small data and calibration requirements and the ability to incorporate changes in climate and landscape properties into the predictive framework. These advantages, along with the ability to disentangle inter-annual and intra-annual variations of low flows offer considerable scope to use this low dimensional modeling approach to inform infrastructure design and water resources policies.

Chapter 4

A topological restricted maximum likelihood (TopREML) approach to regionalize trended runoff signatures in stream network

4.1 Introduction

Regionalizing runoff and streamflow for the purposes of making Predictions in Ungauged Basins (PUB) continues to be one of the major contemporary challenges in hydrology. At global, regional and local scales only a small fraction of catchments are monitored for streamflow [32], and this fraction is at risk of decreasing given the ongoing challenge of maintaining existing gauging stations [217]. Reliable information about local streamflows is essential for the management of water resources, especially in the context of changing climate, ecosystem and demography; and flow prediction uncertainties are bound to propagate and lead to significantly suboptimal design and management decisions [e.g., 207, 214]. Techniques are needed to maximize the use of available data in data scarce regions to accurately predict streamflow, while providing a reliable estimate of the related modeling uncertainty.

There are a number of approaches to predicting runoff in ungauged catchments, including process-based modeling, as presented in Chapter 3, graphical methods based on the construction of isolines [e.g., 28], and statistical approaches. Statistical approaches are often implemented via linear regression, wherein the runoff signature of interest is considered to be an unobservable random variable correlated with observable features of both gauged and ungauged basins (e.g. rainfall, topography). Such linear models are well understood and widely implemented, not only for PUB [see review in 32, p.83] but also across a wide variety of fields in the physical and social sciences [e.g., 156].

4.1.1 Linear Models

Spatial correlation is generally problematic for linear model predictions, including the multiple regression approaches commonly used for regionalization. For example if these models predict a hydrologic outcome y using a matrix X of observed features then the linear model has the form:

$$y = X\tau + \eta \tag{4.1}$$

Here τ is an *a priori* unknown set of weights that represent the influence of each external trend on the hydrological outcome being modeled. The residuals, η, are the observed variation of y that cannot be explained by a linear relation with X. If the residuals are *independent and identically distributed* (iid), the *best linear unbiased predictions* (BLUP) of both y and its uncertainty (i.e. $\mathsf{Var}\,(y)$) can readily be obtained using ordinary least squares (OLS) regression. Unfortunately, residuals are rarely *iid* in hydrological applications due to the spatial organization of hydrological processes around the topology of river channel networks. This organization has the potential to introduce non-random spatial correlations with a structure imposed by the river network. To recover a suitable model in which residuals remain independent requires that the model structure be altered to explicitly account for the spatial and topological correlation in the residuals.

4.1.2 Spatial Correlation Models

There are several techniques available to address spatially correlated data. Within PUB, kriging [64] based geostatistical methods have been widely used [e.g, 108, 97, 193, 194, 208]. In a geostatistical framework, a parametric function is used to model the relationship between distance and covariance in observations. The ensuing semi-variogram is assumed to be homogenous in space, and predictions at a point are computed as a weighted sum of the available observations. The weights are chosen to minimize the variance while meeting a given constraint on the expected value of the prediction. In ordinary kriging for PUB applications, that constraint is simply the average of the streamflow signature as observed in gauged catchments. Ordinary kriging can also be extended as 'universal kriging' to include a linear combination of observable features [167]. Kriging approaches are widely used to predict spatially-distributed point-scale processes like soil properties [e.g., 95] and climatic features [e.g., 94]. Although ordinary kriging has also been used to interpolate runoff [e.g., 108], the theoretical justification for this approach is less robust than for point-scale processes. Runoff is organized around a topological network of stream-channels, and the covariance structure implied for prediction should reflect the higher correlation between streamflow at watersheds that are 'flow connected' (i.e. share one or more subcatchments), compared to unconnected but spatially proximate catchments. Currently, two broad classes of geostatistical methods accommodate this network-aligned correlation structure.

The first suite of methods posits the existence of an underlying point-scale process, which is assumed to have a spatial autocorrelation structure that allows kriging to be applied. Because the runoff point-scale process is only observed as a spatially integrated measure

made at specific gauged locations along an organized network of streams, the spatial autocorrelation structure of the point-scale process cannot itself be observed. Block-kriging approaches [97, 193, 194] infer the semi-variogram of the (unobserved) point-scale so as to best reproduce the observed spatial correlation of the area-integrated runoff at the gauges – a procedure known as regularization. The topology of the network is implicitly accounted for by the fact that nested catchments have overlapping areas, which affects the relation between observed (area integrated) and modeled (point scale) covariances. Yet, complex catchment shapes complicate the regularization of semi-variograms, meaning that the estimation of the point-scale process becomes analytically intractable and requires a trial-and-error approach in most practical applications (e.g., Top-kriging [208]). Top-kriging is an extension of the block-kriging approach that accommodates non-stationary variables and short observation records. Top-kriging provides an improved prediction method for hydrological variables when compared to ordinary kriging or linear regression techniques [131, 231, 53] and was recently extended to account for deterministic trends [132]. Top-kriging represents an important advance for PUB, but it does have a few drawbacks: (i) The regularization process is unintuitive, and requires extensive trial-and-error to determine both the form of a suitable point-scale variogram, and its parameters; (ii) This trial-and-error process is likely to be computationally expensive; (iii) Like all kriging techniques, the estimation of the variogram is challenging when accounting for observable features: the presence of an unknown trend coefficient and variogram leads to an under-determined problem, making consistent estimates for both challenging. Cressie [64] (p. 166) showed that the presence of a trend tends to impose a spatially inhomogeneous, negative bias on the estimated semivariogram. The bias increases quadratically with distance, meaning that estimates of the long-range variance (the *sill*) are strongly impacted by the presence of the trend, leading to an underestimation of the predication uncertainty. This bias, however, only marginally affects the prediction itself.

Geomorphological considerations of the topology of a river network generally focus on the channels, and lead to an intuitive conceptualization that topological interpolation should focus on runoff correlations along flow paths. The second type of approach embraces this topological structure. It does not consider a point-scale runoff generation process, but instead models the hillslope-scale runoff delivery process to the channel network as a unidimensional directed tree [65, 230]. Runoff correlation is expected to decrease with the distance along the stream following a known parametric function. However, unlike Euclidian distances, the stream-wise distance does not have the necessary properties to provide a solvable kriging system. This issue is addressed in Cressie et al. [65] and Ver Hoef and Peterson [229], where streamflow is modeled as a random process represented by a Brownian motion that starts at the trunk of the tree (i.e. the river mouth) moves upstream, bifurcates and evolves independently on each branch. The resulting model only allows spatial dependence with points that are upstream on the river network and provides a positive definite covariance matrix that is estimated through restricted maximum likelihood (REML). Models of this nature have been successfully tested on stream chemistry data [230] and further developed to also allow spatial autocorrelation among random variables on stream segments that do not share flow, with potential applications to the modeling of the concentration of upstream moving species

(e.g., fishes or insects) [229]. While these methods do not account for the streamflow generation process, they avoid the conceptual and prediction uncertainty challenges confronted by kriging techniques.

4.1.3 The TopREML Approach

Inspired by both types of approaches, here I present a method based on the use of linear mixed models to generate a BLUP for hydrological variables on a flow network. Rather than using a kriging estimator, I adopt a Restricted Maximum Likelihood (REML) framework [90, 170, 136] to estimate variance parameters. This reduces the bias on the semivariogram by allowing the variance to be estimated independently from the trend coefficients [64, 136]. This use of a REML framework to estimate a linear mixed effect model on a topological support is termed TopREML. The approach is based on the following conceptual assumptions:

(a) Flow generation and propagation:

Similar to Top-Kriging, runoff is assumed to be generated at a point scale on the landscape, from where it is routed to a channel and measured at a gauge (Figure 4.1 (i)). Runoff observations made at any individual gauge (Figure 4.1 (ii)) can be broken up into a *local* contribution, derived from a never-previously-gauged catchment area, and an *upstream* contribution that was previously observed at upstream gauge(s) along the channel (Figure 4.1 (iii)). TopREML disaggregates all flow contributions into a cascade of local components, as observed at each successive gauge, and uses these characteristics to constrain the covariance structure of runoff and to account for the stronger spatial correlations between flow-connected basins.

(b) Treatment of time:

For the local effects to form a suitable basis for spatial interpolation, variations associated with temporal correlation (e.g. travel time effects) need to be removed. This is achieved by considering time-averaged streamflow data, with the proviso that the time averaging window is much greater than the characteristic catchment and channel response timescales. This treatment of time has several specific consequences. First, TopREML is only suitable for the regionalization of time-averaged and statistically stationary runoff properties (i.e. *runoff signatures*). Stationarity is necessary to ensure that the water balance assumption used to separate local from upstream runoff contributions is valid. However, as a consequence, TopREML cannot be used to interpolate transient signatures, such as those associated with real-time forecasting. Nor can it be used to describe runoff properties that are correlated over time scales larger than the time averaging window. Because of the stationarity assumption applied, all correlation arguments described in this manuscript refer to the *spatial*, and not *temporal*, correlation of the runoff signatures.

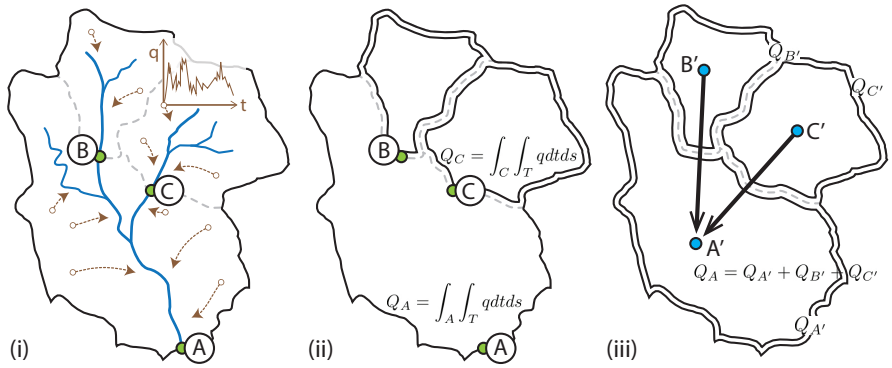


Figure 4.1: Conceptual flow propagation model. (i) Runoff is generated continuously by a spatially distributed point process and drained to the stream network. (ii) When monitored by stream gauges, runoff is spatially integrated over the corresponding catchment and temporally averaged at the chosen observation frequency (e.g., daily streamflow). (iii) The model conceptualizes the catchments as isolated drainage areas (A', B', and C') representing the *local* runoff contribution to each gauge. The flow actually measured at each gauge is the sum of the upstream isolated drainage areas.

(c) Network topology:

Network topology in TopREML also follows a conceptual model that is similar to the model posited by Top-kriging. Topology is conceptualized by area connectivity. That is, flow-connected gauges are characterized by overlapping drainage areas. Unlike Top-kriging, TopREML does not require information about a spatially random point process, but solely relies on information measured at the gauges. It uses the inter-centroidal Euclidian distance between drainage areas of the local flow contributions at each gauge – the *isolated drainage areas* (IDA) – as a distance metric to compute streamflow correlation. The underlying assumption is that runoff signatures of local flow generation regions that are close to each other (in Euclidian space) are more likely to be identical. Although TopREML doesn't require that the characteristics of a point-scale runoff generation process are known in order to support interpolation (a necessary requirement for Top-kriging), the existence of such a point process is consistent with the treatment of spatial correlation in TopREML. To illustrate this consistency, a stylized example relating point-scale runoff generation to the existence of a covariance-structure that relates flow-connected gauges is outlined as an Appendix (Appendix C.1).

4.1.4 Chapter Outline

I first derive the TopREML estimator and its variance for mass conserving (i.e. linearly aggregated) variables, with extensions to some non-conservative variables (Section 4.2). I then apply the approach in two case studies to evaluate its ability to predict mean runoff and

runoff frequency by comparison to other available interpolation techniques: Sections Section 4.3.1 and Section 4.4.1 present leave-one-out cross-validations in Nepal (sparse gauges, significant trends) and Austria (dense gauge network, no observed trends). In both cases, TopREML performed similarly to the best alternative geostatistical method. I then use numerical simulations to illustrate the effect of the two distinguishing features of TopREML: its ability to properly predict runoff using highly nested networks of stream gauges and its ability to properly estimate the prediction variance when accounting for observable features (Sections Section 4.3.2 and Section 4.4.2). Finally, I discuss the limits and delineate the context in which TopREML – and geostatistical methods in general – can successfully be applied to predict streamflow signatures in ungauged basins (Section 4.5).

4.2 Theory

4.2.1 Accounting for spatially correlated residuals

Linear models can be used to make predictions about hydrological variables along a network, provided that the models explicitly address the effects of network structure. A mixed linear model approach provides a suitable framework for this accounting. In this framework, the effects of observable features on the hydrological outcome are assumed to be independent of the network, and retain their influence independently, as so-called 'fixed effects'. The role of spatial structure is assumed to lead to correlation specifically in the residuals η. The residuals are split into two parts: (i) one containing 'random effects' u that exhibit spatial correlation along the flow network and (ii) a remaining, spatially independent, white noise term ε, which does not have any spatial structure. With these assumptions, the mixed linear model is written as:

$$y = \underbrace{X}_{\substack{\text{Trends:}\\ \text{Explanatory}\\ \text{variables}\\ \text{(Nxk)}}} \underbrace{\tau}_{\substack{\text{Coefficients}\\ \text{(kx1)}}} + \underbrace{I_N}_{\substack{\text{Identity}\\ \text{Matrix}\\ \text{(NxN)}}} \underbrace{u}_{\substack{\text{Correlated}\\ \text{random}\\ \text{effects}\\ \text{(Nx1)}}} + \underbrace{\varepsilon}_{\substack{\text{Residuals,}\\ \text{uncorrelated}\\ \text{errors}\\ \text{(Nx1)}}} \tag{4.2}$$

To proceed, I assume that u and ε (and therefore y) are normally distributed with zero mean and are independent from each other. The variance associated with ε is denoted σ^2, the variance of u is assumed to be proportional to σ^2 according to some ratio, ξ, and finally, u is assumed to have a spatial dependence captured by a correlation structure G, which is related to the spatial layout of gauges along the river network and a distance parameter ϕ (the correlation range). Thus, the random effects can be specified as:

$$\begin{bmatrix} u \\ \varepsilon \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \sigma^2 \begin{bmatrix} \xi G(\phi) & 0 \\ 0 & I_N \end{bmatrix} \right) \tag{4.3}$$

To solve this mixed model, five unknowns must be found: σ^2 , ξ, ϕ, the fixed (τ) and random (u) effects. Once τ and u are known, the empirical best linear unbiased prediction (E-BLUP) of y can be made at ungauged locations [136]. The solution strategy adopted here is

to prescribe a parametric form for $G(\phi)$, allowing the covariance structure along the network to be specified, and the likelihood function for the model to be written in terms of *all* five unknowns. Identifying the parameter values that optimize this model thus simultaneously solves for the correlation structure, covariance parameters, fixed and random effects. To proceed with the specification of $G(\phi)$, however, the form of the covariance structure that arises along the network needs to be addressed.

4.2.2 Covariance structure of mass conserving variables

In the linear mixed model framework, the propagation of hydrological variables through the flow network introduces topological effects into the covariance structure of that variable. Firstly, linearly propagated variables, such as annual specific runoff, are discussed. Nonlinearly-propagating variables can in some cases be transformed to allow the linear solutions to be used (as outlined in Section 4.2.5). Consider a set of streamflow gauges monitoring a watershed as illustrated in Figure 4.1 (ii). Because of the nested nature of the river network, the catchment area related to any upstream gauge is entirely included within the area drained by all downstream gauges. To account for the network structure, the catchment at any location along a stream can be subdivided into the *isolated drainage areas* (IDA) that are *monitored for the first time* by an upstream gauge. This is illustrated in Figure 4.1 (iii), and leads to a subdivision into non-overlapping areas, each associated with the most upstream gauge that monitors them. In making this subdivision, it is implicitly assumed that the timescales at which a hydrological variable is propagated in the channel are negligible compared with the timescales on which hillslope effects operate (a generally valid assumption for small to moderately sized watersheds [see 71]). IDA's can be associated with both gauged locations and ungauged locations. In what follows, indices i, j, k, and m are used to refer to gauged sites, while index n refers to ungauged sites where a prediction is to be made.

With these assumptions, observations of y_i made at gauge i can be expressed as a linear combination of contributions from the upstream IDAs:

$$y_i = \sum_{k=i}^{k \in \mathsf{UP}_i} a_k y'_k \tag{4.4}$$

where y'_k is the contribution of the IDA related to gauge k (that is, y_i is equivalent to y'_i only if there are no gauges upstream of gauge i); UP is the set of isolated drainage areas monitored by gauges that are located upstream of i; $a_k = A_k / \sum_{m=i}^{\mathsf{UP}} A_m \leq 1$ is the surface area of the drainage area k normalized by the total watershed area upstream of gauge i. The covariance between observations of y made at different gauges can then be expressed as

$$\begin{aligned}\mathsf{Cov}\left(y_i, y_j\right) &= \mathsf{E}\left[y_i y_j\right] - \mathsf{E}\left[y_i\right]\mathsf{E}\left[y_j\right] \\ &= \sum_{k=i}^{k\in\mathsf{UP}_i}\sum_{m=j}^{m\in\mathsf{UP}_j} a_k a_m \mathsf{E}\left[y'_k y'_m\right] - \left(\sum_{k=i}^{k\in\mathsf{UP}_i} a_k \mathsf{E}\left[y'_k\right]\right)\left(\sum_{m=j}^{m\in\mathsf{UP}_j} a_m \mathsf{E}\left[y'_m\right]\right)\end{aligned}$$

With $\mathsf{E}\left[y'_k y'_m\right] = \mathsf{Cov}\left(y'_k y'_m\right) + \mathsf{E}\left[y'_k\right]\mathsf{E}\left[y'_m\right]$, I have

$$\mathsf{Cov}\left(y_i, y_j\right) = \sum_{k=i}^{k\in\mathsf{UP}_i}\sum_{m=j}^{m\in\mathsf{UP}_j} a_k a_m \mathsf{Cov}\left(y'_k, y'_m\right) \tag{4.5}$$

where $\mathsf{Cov}\left(y'_k, y'_m\right)$ is the covariance between the contributions of sub-catchments k and m. By summing over UP in Equation (4.5) (rather then the complete set of available gauges), the model assumes no correlation between runoff observed at flow-unconnected gauges.

Here I assume that the area-averaged process y' is drawn from a second order stationary random process, and that the covariance between y'_k and y'_m will depend only on the relative position of sub-catchments m and k, given some specified correlation function $\rho(\cdot)$ of the distance c_{km} between the centroids of the two sub catchments [64]. I assume that this function is well approximated by an exponential function $\rho(c_{km}, \phi) = exp(-c/\phi)$. A justification for this assumption, which reproduces the streamflow variances observed in my case studies well (Figure C.1), is derived for strongly idealized conditions in Appendix C.1. Finally, because the observations made at the gauges represent an area-averaged process, the averaging generates a nugget variance σ^2 that is homogenous across observations. The nugget consists of the variance of processes that are spatially correlated over scales smaller than the sub-catchments (see Appendix C.1) and of measurement errors at the gauges.

With this background, the covariance matrix of y can be expressed as

$$\mathsf{Cov}\left(y_i, y_j\right) = \xi\sigma^2 \sum_{k=i}^{k\in\mathsf{UP}_i}\sum_{m=j}^{m\in\mathsf{UP}_j} a_k a_m \rho(c_{km}, \phi) + \sigma^2 = \sigma^2 \cdot (\xi U[A \diamond R]U^T + I_N) \tag{4.6}$$

where $\sigma^2 = \mathsf{Var}\left(y'_k, y'_k\right)$, $U_{i,j} = \mathbf{1}\{j \in \mathsf{UP}_i\}$, $A = aa^T$, and $R_{i,j} = \rho(c_{i,j}, \phi)$. $[\cdot \diamond \cdot]$ denotes the element-by-element matrix multiplication. The matrix G describing the correlation between the random effects in Equation (4.3) is finally

$$G(\phi) = U[A \diamond R(\phi)]U^T \tag{4.7}$$

The topology of the network is described by the matrix U, which ensures that only those catchments that are on the same sub-network (upstream or downstream) of the considered gauge are utilized in the determination of the covariance of y. This spatial constraint comes at the expense of neglecting potential correlations with neighboring catchments that are not flow-connected, and the effects of this tradeoff are investigated in the Monte Carlo experiment described in Section 4.3.2. The effect of spatial proximity is addressed by use of the Euclidian distance between catchment centroids (matrix R), and the effect of scale is accounted for by weighting by the catchment area of the IDAs (matrix A).

4.2.3 REML estimation

The restricted maximum likelihood approach partitions the likelihood of

$$y \sim \mathcal{N}\left(X\tau, \sigma^2(\xi G + I_N)\right)$$

into two parts, one of which is independent of τ [63]. This allows the determination of fixed effects and the variance parameters of the model (here σ^2, ϕ and ξ) to be undertaken separately. The variance parameters are then estimated by maximizing the restricted log likelihood expression [89]

$$\lambda_R(\sigma^2, \phi, \xi) = -\frac{1}{2}\left(\log\det\left(X^T H^{-1} X\right) + \log\det\left(H\right)) + \nu \log\sigma^2 + \frac{1}{\sigma^2} y^T P y\right)$$

where $\det(\cdot)$ is the matrix determinant operator, $\nu = N - k$, $H = I_N + \xi G$, and $P = I_N - WK^{-1}W^T$, $W = [X : I_N]$ and R is the correlation matrix in Equation (4.7), and K is the block matrix:

$$K = \begin{bmatrix} X^T X & X^T \\ X & I_N + \xi^{-1} G^{-1} \end{bmatrix}$$

The REML estimators $\hat{\sigma}^2$ and $\hat{\phi}$ that maximize λ_R can be obtained through numerical optimization.

4.2.4 E-BLUP and prediction variance at ungauged catchments

Once the variance components $\hat{\phi}$ and $\hat{\xi}$ are estimated, the fixed effect coefficients $\hat{\tau}$ and the random effects $\tilde{u}$ can be obtained by solving the linear system [104]:

$$K(\hat{\phi}, \hat{\xi}) \begin{bmatrix} \hat{\tau} \\ \tilde{u} \end{bmatrix} = \begin{bmatrix} Xy \\ y \end{bmatrix} \tag{4.8}$$

The empirical best linear unbiased prediction of $\tilde{y}_n$ at an ungauged site n can be computed by summing the fixed and random effect predictions [136]

$$\tilde{y}_n = x_n^T \hat{\tau} + \tilde{u}_n = x_n \hat{\tau} + g_n^T G^{-1} \tilde{u} \tag{4.9}$$

where x_n is the vector of fixed covariates at ungauged site n, g_n a correlation vector between site n and each gauge; given $\hat{\phi}$, g_n can be readily obtained from the relative position of site n and the gauges in the river network.

The variance of the TopREML prediction error can be expressed as

$$\begin{aligned} \mathsf{Var}\left(\tilde{y}_n - y_n\right) &= \mathsf{Var}\left(x_n^T(\hat{\tau} - \tau) + g_n^T G^{-1}(\tilde{u} - u)\right) \\ &= x_n^T \mathsf{Var}\left(\hat{\tau} - \tau\right) x_n + g_n^T G^{-1} \mathsf{Var}\left(\tilde{u} - u\right) G^{-1} g_n + 2 x_n^T \mathsf{Cov}\left(\tilde{u} - u, \hat{\tau} - \tau\right) G^{-1} g_n \end{aligned} \tag{4.10}$$

The covariance matrix of the error on τ and u in Equation (4.10) can be expressed as a function of the inverted model matrix K [136]:

$$\mathsf{Cov}\begin{pmatrix} \hat{\tau} - \tau \\ \tilde{u} - u \end{pmatrix} = \sigma^2 K^{-1} \tag{4.11}$$

This provides:

$$\mathsf{Var}\left(\tilde{y}_n - y_n\right)_- = \sigma^2 \left(x_n^T K_{11}^{-1} x_n + g_n^T G^{-1} K_{22}^{-1} G^{-1} g_n + 2 x_n^T K_{12}^{-1} G^{-1} g_n\right) \tag{4.12}$$

Where K_{11}^{-1}, K_{22}^{-1}, K_{12}^{-1} are $k \times k$, $N \times N$ and $k \times N$ partitions of the inverted K matrix. If ε is an error that is truly *iid* and does not affect the true value of y_n (e.g., measurement errors), then Equation (4.12) corresponds to the mean square error of the TopREML prediction of y_n. If, by contrast, ε represents random variations of the true value of y_n that are correlated over short distances (and so do not appear correlated in my data), then ε should be included in Equation (4.10) and the prediction variance becomes

$$\mathsf{Var}\left(\tilde{y}_n - y_n\right)_+ = \mathsf{Var}\left(\tilde{y}_n - y_n\right)_- + \sigma^2, \tag{4.13}$$

because ε and u are independent. In reality ε is likely composed of both spatially correlated and *iid* error components and the true variance will be somewhere between these two bounds [136].

4.2.5 Application to non-conservative variables

Unlike mean specific runoff, numerous streamflow signatures (e.g., runoff frequency or descriptors of the recession behavior) are non-conservative and cannot be expressed as linear combinations of their values in upstream sub-catchments. In such conditions the derivations in Section 4.2.2 cannot be applied and the correlation structure in Equation (4.7) will lead to biased REML predictions. The effect of the network structure on streamflow can nonetheless be accounted if the non-linearities can be neglected or eliminated through algebraic transformations.

For instance, runoff frequency λ, as used in Chapter 3, is defined as the probability, on daily timescales, that a gauge will record a positive increment in streamflow. Provided all sub basins are large enough to significantly contribute to streamflow, a runoff pulse at *any* of the upstream sub-basins causes a streamflow increase at the gauge. Therefore runoff frequency does not scale linearly through the river network. It can nonetheless be shown (see Appendix C.2) that if runoff pulses occur independently for each sub-basin, the logarithm of the complement to runoff probability (i.e. $ln(1-\lambda)$) propagates linearly throughout the network enabling the application of TopREML to predict runoff probability at ungauged catchments.

A similar reasoning can be applied to predict recession parameters. For example, the exponential function $Q(t) = Q_0 exp(-k_r t)$ is a widely used approach to model base flow

recession, where $Q(t)$ is the discharge at time t, Q_0 the peak discharge, and k_r the recession constant which can be considered to represent the inverse of the average response time in storage [238]. Because expected values scale linearly, the average response time at a gauge can be modeled as a linear combination of the mean response times of the upstream IDAs. Therefore, although recession constants themselves do not propagate linearly, their value in ungauged basins can be estimated by taking the inverse of TopREML predictions of average response times.

4.2.6 Implementation

TopREML is implemented in R [113], as described in Appendix C.3. The script is provided as a supplement to this manuscript. To run the script, two vector datasets (e.g., ESRI Shapefile) are needed as inputs – one containing the catchments where runoff is available and another containing the basins where predictions are to be made. Catchment polygons and explanatory and predicted variables must be provided as attributes of the vector polygones. The way in which the catchment polygons are nested provides the topology of the stream network. TopREML uses the BFGS algorithm [242] to maximize the restricted log likelihood, though stochastic algorithms are required if a non-differentiable (e.g., spherical) covariance function is selected. The selection of initial values for σ^2, ϕ and ξ is a key user input that may affect the performance of optimization algorithms by causing them to converge to a local extrema. I found that initial values of $[\sigma_0^2, \phi_0, \xi_0] = [\sigma_{LM}^2, \mathsf{E}[c_{km}], 1]$ worked well in my case studies, with σ_{LM}^2 the variance of the OLS residuals of the linear model and $\mathsf{E}[c_{km}]$ the average distance between IDA centroids.

4.3 Methods

4.3.1 Case studies

Observed streamflow data are used to evaluate the ability of TopREML to predict streamflow signatures in ungauged basins. The assessment is based on leave-one-out cross validations, where the tested model is applied to predict runoff at one basin based on observations from all the other basins. After predicting runoff at all available basins in that manner, the model is evaluated based on its mean absolute prediction error. Streamflow variables from 57 catchments in Upper Austria [209] and 52 catchments in Nepal [66] are used in two separate leave-one-out analyses. The location of the gauges is shown in Figure 4.2, and Table 4.1 provides a summary of relevant catchment characteristics. Further details on the datasets are provided in Skøien et al. [209] for Austria and Chapter 3 in Nepal. The two regions differ significantly with respect to gauge density (high in Austria and low in Nepal) and in the nature of the runoff signature and observable features. The Nepalese datasets provides specific runoff and wet season runoff frequency, as well as gauge elevation and bias-adjusted annual rainfall derived from the Tropical Rainfall Measurement Mission 3B42v7 dataset [152].

Gauge elevation and annual rainfall are used as observable features for specific runoff [56]. The Austrian dataset was directly taken from the *rtop* package [209], where mean summer runoff observations are provided to demonstrate Top-kriging. The Austrian dataset did not contain additional observable features and previous studies have found spatial proximity to be a significantly better predictor of runoff than catchment attributes in Austria [147].

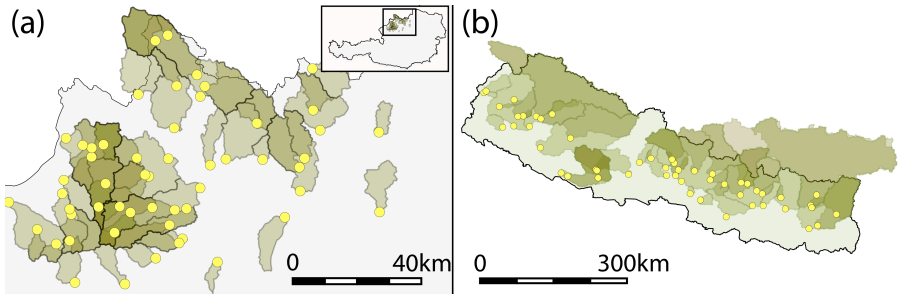


Figure 4.2: Location of the gauges and related catchments included in the cross validation analyses in Upper Austria (a) and Nepal (b). Coloring is semi-transparent to emphasize overlapping catchment areas.

	N	Q	λ	A	c	Dpt	P_y	z_g
Nepal	52	1660	0.42	2121	13.9	10	1683	320
		(1062, 2228)	(0.40, 0.46)	(513, 5267)	(9.2, 25.2)		(1482, 1909)	(507, 750)
Austria	57	0.68		68	4.5	8		
		(0.42,1.43)		(44, 136)	(3.9, 6.3)			

Table 4.1: N is the number of catchments; Q the specific runoff $[mm/y]$ in Nepal and the mean summer streamflow $[m^3/s]$ in Austria; λ is the rainy season runoff frequency (d^{-1}) in Nepal; A the catchment area in km^2; c the distance in km between the centroids of isolated drainage areas; Dpt the depth of the stream network graph (i.e. the maximum number of flow-connected gauges); P_y the annual rainfall in mm given by TRMM over Nepal and adjusted according to [152]; z_g is the gauge elevation in meters above sea level. Median values are provided with 25th and $75th$ quantiles in parenthesis.

The predictive ability of TopREML was evaluated on (a) specific annual runoff in Nepal, (b) wet season runoff frequency in Nepal and (c) average summer streamflow in Austria. The performance of TopREML (TR) was compared to five other widely used regionalization methods: sample mean (LM_0), linear regression (LM), universal kriging (UK) and Top-kriging (TK). As shown in Table 4.2, these methods cover a wide spectrum of incrementally specific assumptions and comparing them provides an assessment of the value added by increased model complexity for regionalization of these streamflow parameters. Code to implement all four methods is readily available in R, with dedicated packages available for Top-kriging – *rtop* – and universal kriging – *gstat* [171].

4.3.2 Numerical Simulations

Network Effects Conventional geostatistical methods predict runoff by weighing observations from surrounding basins based on their geographic distance. TopREML also incorporates the topology of the stream network by including or excluding basins based on their flow-connectedness. This adds topological information to the determination of the covariance structure of runoff, at the expense of discarding information that could be derived from cor-

	Explanatory Variables	Spatial Covariance	Network Topology	Unbiased Variance
Sample mean				
Linear regression	X			
Universal kriging	X	X		
Top-kriging	X	X	X	
TopREML	X	X	X	X

Table 4.2: Taxonomy of the compared regionalization approaches.

relations between spatially proximate regions that are not connected to the gauge of interest by a flow path. Assessing the net benefits of accounting for network effects requires being able to control the topology of the network, and thus requires numerical simulations. A series of Monte Carlo experiments as described in Figure 4.3 were run to simulate network complexity by varying the number of flow-connected basins that are within (N_{inner}) and beyond (N_{outer}) the predefined spatial auto-correlation range of the randomly generated runoff. A non-topological geostatistical method like universal kriging would include all basins within and exclude all basins beyond the spatial auto-correlation range. I expect TopREML to outperform universal kriging when the number of flow-connected basins beyond the autocorrelation range increases and the number of connected basins within the autocorrelation range decreases.

Variance Estimation and Observable Features. A key advantage of the Reduced Maximum Likelihood framework is its ability to avoid the downward bias in the covariance function that affects kriging-based methods (including Top-kriging) when external trend coefficients are simultaneously estimated. This bias particularly affects the prediction of the variance. Again, empirical cross validation analysis does not allow an assessment of this bias, because the observation datasets used contained only one observation per location. Numerical simulations, however, allow many realizations of the underlying stochastic process to be made at each location, and thus allow the prediction variance to be compared with the numerical variance. I evaluate TopREML's ability to predict variances (and therefore evaluate prediction uncertainties) at ungauged locations using the Monte Carlo procedure on the synthetic catchments described in Figure 4.3. I construct the observed prediction uncertainty by taking the standard deviation of the prediction errors across all 1000 Monte Carlo runs and compare it to the square root of the median predicted variance. The external trend is omitted from the model specification (i.e. it is *not observed*) in a first experiment, and explicitly included in the model in the second experiment. I compare TopREML and Top-kriging based on their ability to model prediction variance. I expect TopREML to provide a better estimate of the variance than Top-kriging when accounting for observable features. Because the trend is spatially correlated, omitting it in the model specification adds a significant spatially correlated component to the error and Equation (4.13) should be used to predict the variance. Conversely, including a trend in the model will cause the remaining

error to mostly consists of (spatially uncorrelated) residuals so in this case Equation (4.12) is used.

4.4 Results

4.4.1 Case Studies

Basin-level predictions of the considered signatures are presented in Figure 4.4 for the three cross validation analyses described in Section 4.3.1. Figure 4.4 also provides box plots summarizing the distribution of the ensuing cross validation errors. In the three analyses, the prediction errors related to TopREML were comparable to the best alternate method: a linear model for annual specific runoff (Nepal) and Top-kriging for runoff frequency (Nepal)) and summer runoff (Austria).

Figure 4.4 (a) presents results for annual streamflow in Nepal and shows that observable features play a significant role in the prediction of runoff. The linear model showed a highly significant effect of annual precipitation ($\hat{\tau}^{(LM)}_{\mathsf{yearlyPrecip}} = 0.99$, t-stat: 9.1) a moderately significant effect of altitude ($\hat{\tau}^{(LM)}_{\mathsf{meanElev}} = 0.39$, t-stat: 2.5) and an overall fit of $R^2 = 0.63$. The positive sign of the altitude coefficient can be attributed to the effects of glacial melt on runoff, which are more significant at higher altitudes, while the average effect of evapotranspiration explains the negative and noisy intercept of -313 mm/y. While including rainfall and altitude in the model decreased the median absolute error by 43% (LM to LM_0), further increasing the complexity of the model by allowing for spatial (UK) and topological effects (TK and TR) did not improve the predictive performance: residuals from the linear regression appeared to be correlated at a range shorter than the distance between the gauges in Nepal. Indeed, fitting the empirical semivariograms with exponential functions revealed spatial correlation ranges that were on the order of the mean distance between IDA centroids for annual streamflow (21.6 km), and significantly below that distance (7.0 km) for the regression residuals. Nonetheless, the lack of parsimony of TopREML did not appear to affect its predictive performance, which almost perfectly reproduced the performance of the linear model – the most parsimonious method.

In contrast, the analysis revealed significant spatial effects for both runoff frequency in Nepal, which has a much larger spatial correlation range than annual streamflow (426 km – presumably set by meteorology and the correlation range of storm events, as asserted in Chapter 2), and summer runoff in Austria, which has a range of 19.1km but is sampled by a much higher density of streamflow gauges than in Nepal. Allowing for spatial correlation in the residuals (UK) decreased the median absolute error by 11% compared to the linear model (LM) for runoff frequency in Nepal and 31% for summer runoff in Austria. Accounting for topological effects further reduced errors by 33% (runoff frequency) and 40% (summer runoff) for both TopREML and Top-kriging methods.

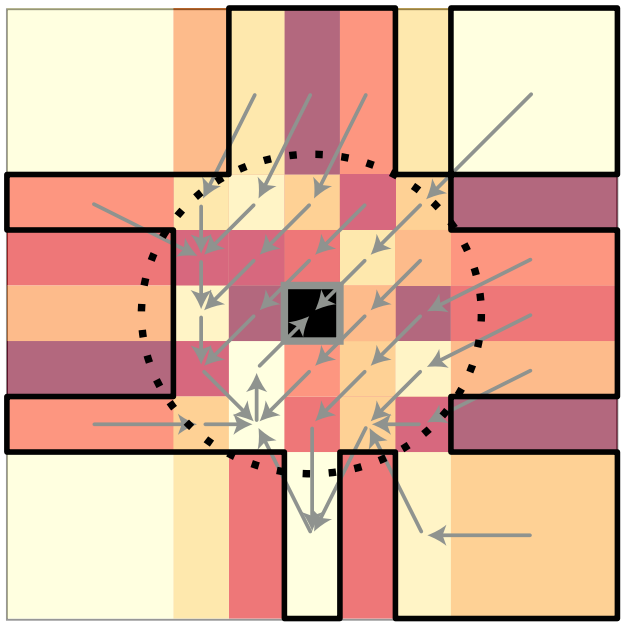


Figure 4.3: Monte Carlo generation procedure: (i) a spatially correlated gaussian field with an exponential covariance function (mean=30 , partial sill=8, nugget=2, range=3) is generated along a 7×7 irregular grid. The central pixel (in black) represents the downstream-most catchment, where runoff is to be predicted. Among the remaining pixels, 24 *inner* isolated drainage areas (IDA) are within a radius of one spatial correlation range (dashed circle) of the central pixel, and 24 *outer* pixels are beyond that radius. (ii) A predefined number of *inner* and *outer* pixels are randomly selected as part of the set of catchments that are flow-connected to the central pixel. In the figure, all 24 *inner* pixels and 12 *outer* pixels are selected and form the flow catchment outlined with a thick black line. (iii) A tree graph is randomly generated (grey arrows) with its trunk at the prediction pixel and branches passing through all the flow connected pixels. The random field generated in step one is aggregated along the tree by summing the value of all lower order branches at each confluence. (iv) A new spatially correlated field (mean=1, partial sill=0.15, nugget=0, range=0.5) is generated at each pixel – that is the observed trend. The trend is multiplied by a predefined trend coefficient (τ=10) and added to the aggregated runoff at each pixel – that is the observed runoff. (v) Based on the observed runoff and (if applicable) trend at the 48 non-central pixels, TopREML and the compared baseline method (top kriging or universal kriging) are used to predict runoff at the central pixel. Prediction errors are recorded and the procedure repeated 1000 times to get the mean and variance of the errors.

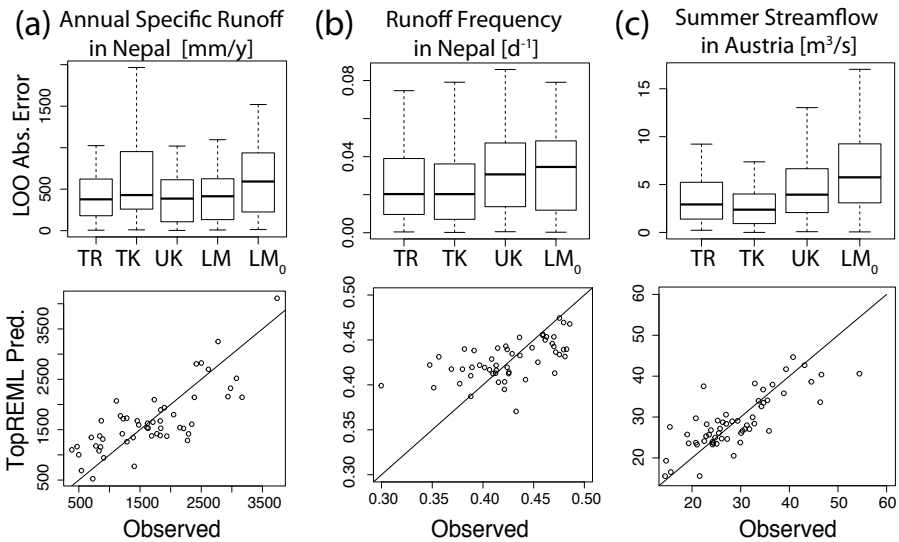


Figure 4.4: Results of the comparative cross validation analyses of (a) specific runoff and (b) wet season runoff frequency in Nepal, and (c) mean summer streamflow in Austria. *First row:* Box plots with the quartiles and 95% confidence intervals around the median of leave-one-out (LOO) absolute prediction errors. Compared models are TopREML (TR), Top-kriging (TK), universal kriging (UK), linear regression models (LM) and the sample mean (LM_0). Note that without observable trends ((b) and (c)), LM and LM_0 are equivalent. *Second row*: Catchment level performance of TopREML. Signatures predicted by TopREML for each catchment in the leave-one-out crossvalidation analysis are plotted against the corresponding observed signature. Diagonal lines (x=y) representing perfect fit are also displayed for indicative purposes.

4.4.2 Numerical Simulation

Results from the Monte Carlo analysis are presented in Figure 4.5, showing the outcomes of the two numerical experiments described in Section 4.3.2.

Figure 4.5 (a) and (b) shows the effect of network complexity on the performance of TopREML relative to the baseline performance of universal kriging. This effect is measured as the difference in the relative errors of the two methods as a function of N_{outer}, the ratio of basins *beyond* the spatial correlation range of runoff that *are* flow-connected, and N_{inner}, the ratio of basins *within* range that are *not* flow-connected. The effect is expected to increase with N_{outer} and decrease with N_{inner}, reaching zero when 100% of observed basins lie within the spatial correlation range and 0% of the basins beyond the range are flow-connected. In that case (not shown in the figure), TopREML and universal kriging perform similarly and the mean difference in the relative error of the two methods is zero. Figure 4.5 (a) shows that the relative performance of TopREML improves with the number of flow-connected catchments that are located beyond the spatial correlation range, and which are therefore not properly accounted for by universal kriging. Conversely, Figure 4.5 (b) shows that

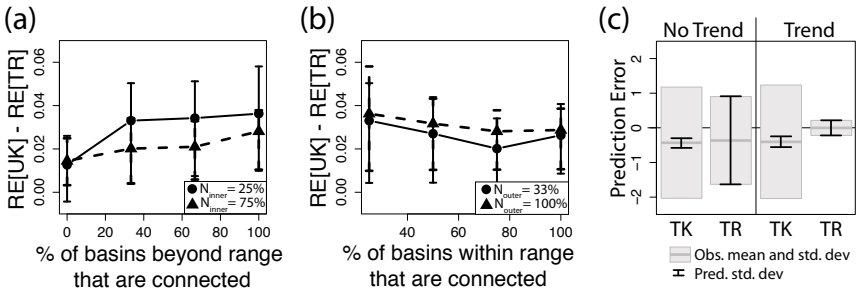


Figure 4.5: Results of the Monte Carlo experiments. Subfigures (a) and (b) display the effect of network complexity on the performance of TopREML relative to universal kriging. Network complexity is given as the ratio of basins *beyond* (N_{outer}) and *within* (N_{inner}) the spatial correlation range that are flow-connected – minimum network complexity is modeled when *no* basins beyond and *all* basins within the range are flow-connected. Relative performance is computed at each Monte Carlo run as the difference in relative prediction errors between universal kriging and TopREML (i.e. $\mathsf{RE[UK]} - \mathsf{RE[TR]}$ on subfigures (a) and (b)). The graphs display the expectation and standard deviation of that difference over the 1000 Monte Carlo runs. Subfigure (c) presents the observed (grey boxes) and predicted (black error bars) standard deviation on the prediction errors for top kriging (TK) and TopREML (TR). Note that the slight downward biases that appear on the graph remain below 1% of the expected value of the predicted outcome.

the relative performance of TopREML decreases with decreasing network effects within the spatial correlation range. A linear regression of the relative performance of TopREML against N_{outer} and N_{inner} showed that both trends are significant and in the expected direction. However, the positive coefficient associated to N_{outer} (9.1, t-stat: 11.9) is larger in absolute value and more statistically significant than the negative coefficient associated to N_{inner} (-2.6, t-stat: -2.6), which suggests that the benefits of including distant flow-connected basins outweigh the costs of discarding nearby (but unconnected) IDAs.

In Figure 4.5 (c), the Monte Carlo analysis showed that model uncertainty is well predicted by TopREML and strongly underestimated by Top-kriging, both with and without considering an external trend. Including a trend in the model reduces the prediction variance of TopREML – this effect is expected because the variance explained by the trend is no longer included in the modeling error ε. The decrease in the prediction variance is well modeled by TopREML, which predicts the observe model uncertainty almost exactly.

4.5 Discussion

4.5.1 Performance of TopREML

Cross validation outcomes suggest that TopREML is an attractive operational tool for predicting streamflow in ungauged basins. The method performs as well as the best alternative approach in the prediction of the considered runoff signatures in Nepal and Austria, and significantly outperforms Top-kriging in the prediction of modeling uncertainties in the numerical analysis.

Two distinguishing features of TopREML are responsible for these encouraging results. First, TopREML incorporates the topology of the stream network by restricting correlations to runoff observed at flow-connected catchments. This allows TopREML to explicitly model the higher correlation in streamflow anticipated along channels, but comes at the expense of discarding correlations with neighboring, but not flow-connected catchments. Such correlations can, for instance, be driven by large scale weather patterns. This tradeoff was investigated in a Monte Carlo analysis showing that modeling performance increases more rapidly when including distant flow-connected basins (slope in Figure 4.5 (a)), than it decreases when discarding nearby unconnected basins (slope in Figure 4.5 (b)). Further, empirical correlograms of Austrian summer runoff (Figure 4.6) reveal significantly lower and shorter-ranged spatial correlations when basins are not flow-connected. Both results suggest that the benefit of accounting for network effects on correlations outweighs the cost of losing some information on the correlation between unconnected basins. Second, the Restricted Maximum Likelihood framework provides an unbiased estimation of variance parameters, even when accounting for observable features. This allows TopREML to accurately predict modeling uncertainties even for highly trended and autocorrelated runoff signatures, as visible in the Monte Carlo analysis presented on Figure 4.5 (c). By contrast, the expected downward bias in the kriging estimation of partial sills [64] is clearly visible in the underestimation of prediction uncertainties by the Top-kriging method.

TopREML also has considerably lower computational requirements than Top-Kriging, both in terms of input data and optimization complexity. Unlike Top-kriging, where watershed polygons are necessary inputs for the regularization procedure, vectors are not fundamentally indispensable for TopREML. Indeed, TopREML does not rely on a distributed point process but assumes homogenous IDAs. It follows that its only fundamental data requirement is a table (i.e. a data.frame) of IDAs displaying the observed regionalization variable and the area, centroid coordinates and network position (i.e. own ID and downstream ID) of the IDA. When considering runtime, both methods rely on numerical optimization, but Top-Kriging uses it to back-calculate the point semi-variogram in its regularization procedure. This may substantially increase the dimensionality of the optimization task, depending on the grid resolution chosen for the discretization of the catchment areas, which in turn has a highly significant effect on prediction performances [208]. By contrast, the dimensionality of the optimization in TopREML is driven by the number of catchments, not an arbitrary grid. More importantly, TopREML admits a well- defined objective function, the

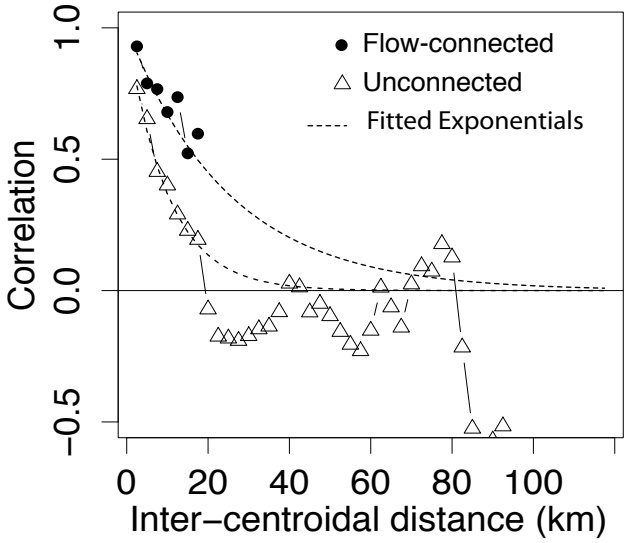


Figure 4.6: Empirical correlograms of the mean specific summer flow recorded at the 57 gauges of the Austrian dataset. Distance has a different effect on the correlation between flow-connected (black circles) and flow-unconnected (white triangles) gauges. Both correlograms are well fitted by an exponential function but the spatial correlation range doubles when gauges are flow connected. Both empirical correlograms are constructed using 5km bins.

restricted likelihood, that is differentiable if the selected variogram function is differentiable. This allows gradient optimization methods to be used, which are much less computationally intensive than the stochastic algorithm required by Top-kriging. The resampling analysis shown in Appendix C.3 suggests that TopREML reduces the computation runtime by an order of magnitude, relative to the implementation of Top-kriging in the *rtop* package, for comparable prediction performances.

Despite these encouraging results, TopREML is subject to stringent linearity assumptions on the nature of the regionalized runoff signature. The predicted variable should aggregate linearly both on hillslope surfaces and at channel junctions that are subject to mass conservation. This limitation also affects block-kriging aproaches, as pointed out by Skøien et al. [208], who suggest that Top-kriging can still be applied, *in an approximate way* on non-conservative variables. Here I assert that hydrologic arguments can be used to convert some non-conservative variables into linearly aggregating processes using simple algebraic transformations. This theoretically more robust approach was here successfully tested in a cross-validation analysis of runoff frequency in Nepal.

4.5.2 Model selection

The regionalization methods assessed in the cross validation analysis range from simple linear regressions with strong independence assumptions, to complex geostatistical methods

that allow for both spatial and topological correlations. Results indicate that while complex methods perform best in general, there seems to be a threshold, beyond which increasing the complexity of the statistical method does not significantly improve the prediction performance: while a linear model is better than a simple average for the prediction of annual streamflow in Nepal (Figure 4.4 (a)), accounting for spatial (UK) and topological (TR) correlation does not further improve predictions. In that situation, parsimony prescribes selecting the least complex of the best performing methods.

Under these conditions, the selection of the optimal method is driven by the interplay between the layout of the gauges and the spatial correlation range of the considered runoff signature. A dense network of flow gauges is necessary for geostatistical methods to properly estimate the semivariogram and improve on predictions from linear regressions – the case studies suggest that the mean distance between the gauges must be on the order of half the spatial correlation range of the runoff signature. Sparser gauge densities do not allow geostatistical methods to capture spatial correlations and their prediction is effectively driven by the deterministic components of the model, i.e. the intercept and (when available) observable features.

An interesting tradeoff arises if observable features are themselves spatially correlated and explain a significant part of the spatial correlation of the predicted variable. Including these observable features in the model reduces the correlation scale of the residuals, possibly crossing the threshold below which geostatistics are not the most parsimonious approach. In Nepal, controlling for rainfall reduced the spatial correlation range of annual streamflow from 21.6 km to 7 km – well below the mean distance between the gauges (13.9 km). In that case there is a tradeoff between relying on observable features or variance information to make a prediction, and parsimony and stationarity considerations come into play when selecting the regionalization model. For instance, while parsimony generally prescribes the use of observable features, a climate may be less stationary – and therefore a less reliable external trend – than embedded geology or geomorphology.

In general, geostatistical approaches improve on the prediction of ungauged basins if the distance between the stream gauges is significantly smaller than the spatial correlation scale of runoff. Favorable areas are characterized by high drainage densities or localized rainfall, in addition to a high density of streamflow gauges. All three variables are highly heterogeneously distributed at a global scale, as seen on Figure 4.7. The multiplicity of local settings likely explains the large diversity of existing regionalization methods and suggests that the selection of the optimal regionalization approach has to be made locally.

Lastly, the decreasing returns to improvements in the complexity of the model also suggest that the performance of statistical methods for PUB is ultimately bounded by the spatial heterogeneity of runoff generating processes. Statistical methods resolve parts of that heterogeneity using the spatial distribution of observable features (linear regressions) and/or based on the analysis of the variance of a sample of the predicted variable (geostatistics). Yet very important parts of the hydrological activity related to storage and flow path characteristics take place underground: they cannot be observed and included in the statistical models [99]. This residual spatial heterogeneity can utlimately only be resolved through a

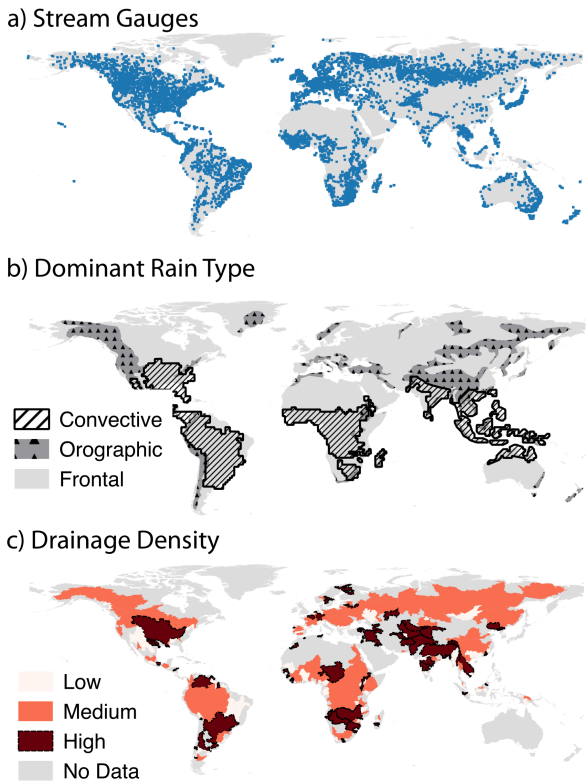


Figure 4.7: Global distribution of factors affecting model selection. (a) Spatial repartition of the 8540 stream gauges indexed by the Global Runoff Data Center [92]. (b) Dominant rainfall type: orographic rainfall are assumed to occur in mountains, as defined by the United Nations Environment Programme [223], and have a typical range of of 1-10km [14]. Convective rainfall are assumed dominant in region with a high frequency of lighting strikes ($\geq 10/km^2yr^{-1}$) as recorded by the TRMM satellite [161] and have a typical scale of 10-100km [34, 210]. Frontal precipitations are assumed dominant in the remaining regions and have a typical scale in excess of 100km [34, 243]. (c) Drainage density is estimated based on the DEM-based Hydro1k dataset [77], using 154 large basins [119] as units of analysis. Drainage densities are displayed in three classes: low ($0.01 - 0.025km^{-1}$), medium ($0.025 - 0.027km^{-1}$) and high ($> 0.027km^{-1}$).

better understanding of the particular catchment processes governing runoff in the considered region. Approaches coupling statistical regionalization with process based models that assimilate both a conceptual understanding of catchment scale processes and the random nature of runoff (e.g., Chapter 3 and [38, 195]) are particularly promising.

4.6 Conclusion

I introduced TopREML as a method to predict runoff signatures in ungauged basins. The approach takes into account the spatially correlated nature of runoff and the nested character of streamflow networks. Unlike kriging approaches, the restricted maximum likelihood (REML) estimators provide the best linear unbiased predictor (BLUP) of both the predicted variable and the associated prediction uncertainty, even when incorporating observable features in the model.

The method was successfully tested in cross validation analyses on mass conserving (mean streamflow) and non-conservative (runoff frequency) runoff signatures in Nepal (sparsely gauged) and Austria (densely gauged), where it matched the performance of the best alternative method: Top-kriging in Austria and linear regression in Nepal. TopREML outperformed Top-kriging in the prediction of uncertainty in Monte Carlo simulations and its performance is robust to the inclusion of observable features.

TopREML's ability to combine deterministic (observable features) and stochastic (covariance) information to generate a BLUP makes it a particularly versatile method that can readily be applied in densely gauged basins, where it takes advantage of spatial covariance information, as well as data-scarce regions, where it can rely on covariates with spatial distributions that are increasingly observable thanks to remote sensing technology. This flexibility, along with its ability to provide a reliable estimate of the prediction uncertainty, offer considerable scope to use this computationally inexpensive method for practical PUB applications.

Chapter 5

Estimating the Elasticities of Distributed Power Generation in Nepal

5.1 Introduction

In a centralized power generation paradigm, electricity is generated at cost-optimal sites and transported to high demand locations through leaky transmission lines. In a country like Nepal, with a complex topography and a sparse and unevenly distributed population located far from hydropower sites, this leads to large transmission losses (34% of the generated electricity [220]) , frequent blackouts and high capital costs for the extension of the electricity grid. To avoid these costs, micro run-of-river hydropower infrastructure can be commissioned to supply isolated and locally owned micro-grids in mountainous areas. Unlike centralized power systems, isolated micro-grids can neither store nor export excess energy, and power generation has to match household electricity demand at the local level. In that context, knowledge of the responsiveness of electricity consumption to changes in its price and, in turn, the responsiveness of electricity costs to the changes in the capacity of the infrastructure, is of crucial to design economically sustainable infrastructure. Yet anecdotal evidences point toward poor design among the causes for the high failure rate of rural micro-hydropower infrastructure [128, 151], which suggests the need for an improved way to evaluate these two effects in data-scarce developing countries. This research aims to contribute to engineering practice by providing an econometrically robust method to estimate the price elasticities of supply and demand, that is the ratios between relative changes in capacity (respectively consumption) and relative changes in price, for micro-hydropower electricity based on a simple survey of existing infrastructure.

Demand The only recent study assessing electricity demand in Nepal considers aggregate energy consumption using national level data [173]. This overlooks rural unmetered

consumers, which are the majority of Nepalese households. As a result, local demand and willingness to pay for new micro-hydropower projects are currently typically evaluated based on the observed consumption of substitute power sources (e.g., batteries, household diesel generators, etc) or stated preferences [125, 84, 124]. These methods are subject to hypothetical biases because the (still inexisting) infrastructure will affect the price and quality of the consumed electricity. To date, there is no widely accepted general theory of respondent behavior to characterize and control hypothetical bias [140], and an alternate approach is required to estimate the elasticity of household electricity demand. Thanks to recent efforts to privatize electricity markets, substantial research has been devoted to assess how household respond to price changes using residential consumption data from power utilities (see, e.g., [80, 106] for a review). Yet few studies (if any) consider distributed generation in developing countries, which differs from centralized grids on three key points:

1. Costs and operational challenges often prevent the installation of household connection meters [189, 48]. As a result, off-grid micro-hydropower schemes in Nepal are typically operated on capacity-based tariffs [85, 87], whereby households pay a fixed fee per unit of electric capacity (e.g. 70 Nepalese Rupees per month for a 100W inlet [122]). The absence of connection meters does not allow households to be billed based on the energy consumed [18]. It follows that electricity consumption choices represent long-term decisions and dynamic (short-term) effects related to, e.g., income or seasonality cannot be observed. Indeed unmetered consumption is driven by the ownership of electrical appliances and limited by an external technical device (a current-limiting fuse): both delay the effect of exogenous shocks on the household's decision to change their level of consumption [114]. The absence of meters also sets the marginal cost of connection time to zero. Households consume more in the absence of meters and elasticity values of *average* consumption obtained from metered data differ from the elasticity of *peak* demand that is relevant for unmetered connections.

2. Second, available evidence suggests that household electricity consumption differs significantly between micro-hydropower- and grid-connected communities, even when controlling for remoteness and population size. In line with Dias-Bandaranaike and Munasinghe [68], this suggests a significant impact of electricity quality on demand. Even in Nepal, where grid blackouts are notorious, micro-hydropower mini grids are usually more prone to outages and voltage fluctuation than the grid because of their small size and undiversified power source [226]. Budhathoki [45] and Ghimire [88] report anecdotal evidence of households switching from micro-hydropower to the grid for power quality reasons. It follows that notwithstanding issues related to the absence of meters, demand curves estimated for grid-connected households cannot directly be attributed to mini-grid users.

3. Finally, existing studies on the price elasticity of electricity demand in developing countries (e.g.,[35, 83, 202]) assume exogenous pricing. One might think that the issue of simultaneity would not matter much in the electricity sector, because prices are regu-

lated and considered to be determined exogenously. This assumption is arguable for grid-connected consumers, where consumers subject to block tariffs can choose the marginal price that would be applied by deciding how much electricity they consume [179, 24]. Price exogeneity is, however, very unlikely for local community-owned infrastructure, where consuming households have a significant influence on price via, e.g., village electrification committees [100].

While the first two points simply suggest that metered connection data cannot be taken as proxies to model the consumption of unmetered household, the last point gives rise to a substantial identification challenge. Price is simultaneously determined by demand and other endogenous factors, notably infrastructure costs, which causal effects are difficult to disentangle. In the absence of experimental data, the causality of interest can be identified using an *instrumental variable*. The chosen instrument should be valid, i.e. correlated to the cause of interest (here electricity price), and exogenous, i.e. not directly correlated to the outcome of interest (here electricity consumption) – the so-called *exclusion restriction*. Instruments based on the components of the price schedule [144] or on non-linearities in the block-tariffs [179] have been applied to resolve this simultaneity issue for residential electricity demand in the US. Yet these methods cannot be readily transferred to unmetered mini grids because the required detailed information on individual household consumption and pricing structure are generally unavailable. In another context, similar identification concerns have also been addressed to resolve endogenous infrastructure placement in the evaluation of the impact of electrification on economic development. To instrument for electricity access, Dinkelman [69] use terrain slope and its effect on the placement of transmission lines, Duflo and Pande [75] exploit non-monotonic relations between slope and dam placement and Rud [190] uses variations in groundwater availability and its effect on the required pumping energy. Unfortunately, these instruments cannot be used in the context of rural Nepal because they are either irrelevant[1], or invalid[2]. Here I build on Lipscomb et al. [139] and use the size of the upstream catchment as instrument for infrastructure cost. I expect the instrument to be valid because, unlike in centralized grids, power is generated locally: infrastructure costs and electricity price reflect local site conditions. I also expect the instrument to be exogenous because the area of the upstream catchment is unlikely correlated to electricity demand other than through its effect on electricity price.

Supply To estimate the effect of the capacity of micro-hydropower infrastructure on their cost, existing studies typically adopt an engineering approach, where total costs are disaggregated into the basic civil, mechanical and electrical components of the plant; the disaggregated costs are then related to local site conditions through standardized cost functions for each individual component [4, 11, 205, 206, 96]. The validity of these estimates hinges on the availability of detailed data on the cost structure and site conditions of a large sample of schemes. This approach is hardly applicable to Nepal, where data is scarce due to the low

[1] There are no dam (as in [75]) and negligible groundwater irrigation (as in [190])

[2] Local topography (as in [69]) likely affects power consumption via agricultural productivity.

accessibility of most sites. As a consequence, existing reviews [e.g., 67, 226] are restricted to a limited number of cases (typically less than 10) with questionable external validity. Unlike existing studies, I explicitly model the equilibrium of the market for decentralized micro-hydropower electricity and use an instrumental variable to estimate the price-elasticity of supply based on the total costs, which are available for a large sample of micro-hydropower plants in Nepal. I also restrict the considered site conditions to a subset of parameters that can be easily observed remotely: local topography, upstream catchment area and distance to nearest road.

The purpose of this chapter is to develop econometrically robust and transferable methods to estimate price elasticities using cross-sectional infrastructure survey data. This will allow constructing the supply and demand curves that are relevant to distributed generation in Nepal, without requiring detailed surveys on local households and site conditions. Section 5.2 presents the available dataset and discusses the empirical strategy used to address the expected simultaneity bias, with regression results presented in Section 5.3. The research contributes to the existing literature in three ways, as discussed in Section 5.4. First, it adds to existing cost estimation techniques for micro-hydropower in the engineering literature [e.g., 4, 11, 205, 206] by providing a robust methodology for data-scarce settings. Second, it adds to the existing literature on the determinants of household electricity demand in developing countries [e.g., 36, 145, 168, 239, 173] by considering the case of decentralized micro-grids in rural Nepal. Finally, this paper contributes to the methodological discussion related to the search for a valid instrument in infrastructure impact studies [e.g., 69, 139, 75, 190].

5.2 Empirical strategy

5.2.1 Data

Electricity consumption data are challenging to obtain because most isolated micro-hydropower schemes in Nepal are unmetered and often unreported, as there are no legal licensing requirements for hydropower plants below 100kW. In addition, unit prices are difficult to obtain from expenditure surveys of grid-connected households because subsidized tariffs are often not representative of the willingness to pay of consumers [36].

Here I use a cross sectional dataset of all micro-hydropower schemes between 1 and 100kW that were subsidized by the Alternative energy Promotion Center (AEPC) of the Government of Nepal and built between 2008 and 2011. Data on plant capacity (kW), total construction costs (C), construction subsidies (S) and the number of supplied households (HH) have been transcribed from the Renewable Energy Data Book (REDB) published by the AEPC [7, 9]. Costs and subsidies are given in Nepalese Rupees ($1\mathsf{USD} \approx 100\mathsf{NRp}$) and normalized to the 2005 constant term. Among the 348 schemes of the dataset, 121 had information on all considered attributes and were included in the analysis. After geocoding each scheme at

the ward level[3], I estimated remoteness (R) as the distance along known footpaths[4] to the nearest motorable road recorded in the gROADS dataset [162]. Unfortunately, the REDB dataset does not indicate the exact location of the scheme within the ward. To recover this information I use a digital elevation model [225] and the simplified algorithm described in Figure 5.1 to identify the river segment most likely to hold a micro-hydropower plant within the territory of each ward. I then estimate the area A of the contributing upstream watershed and use bias-adjusted TRMM precipitation [152] to provide *precip*, the mean annual contributing rainfall. Finally, I match the REDB dataset with VDC level[5] socio-economic community data including $pBiz$, the ratio of households owning a non-agricultural business, which were obtained from the 2001 Census of Nepal.

Two key assumptions are necessary to extract household electricity consumption and unit price from the REDB dataset. First, I assume that the community-owned micro-hydropower schemes in the REDB sample are financially sustainable and the net costs are exactly covered by the user fees. While anecdotal evidence suggests that a large number of micro-hydropower are not financially sustainable, precisely related to the challenges in estimating local demand [e.g., 67], all schemes recorded in the REDB dataset benefitted from subsidies and private loans, which entails some level of due diligence. The AEPC subsidy requirements includes showing evidence of a positive net present value of the infrastructure over 15 years, assuming a 4% discount rate [60]. The theoretical basis of this assumption is further discussed in the determination of the structural equations in Section 5.2.2. The second assumption states that the micro-hydropower schemes are designed to exactly cover the peak consumption of the households, which I assume constant through time. This neglects the dynamic effects of income on electricity consumption. Under these conditions, average household electricity demand KW_{HH} (in kW) can be computed by normalizing the capacity of each scheme by the number of connected households.

$$kW_{HH} = \frac{kW}{HH}$$

Full cost recovery pricing allows the unit price P_s collected by the utility to be exactly equivalent to the unit cost of supplying electricity. This was computed for each plant by normalizing total construction costs by the infrastructure size

$$P_s = \frac{C}{kW}$$

The unit costs actually faced by the households are finally obtained by accounting for construction subsidies:

$$P_{HH} = P_s - \frac{S}{kW}$$

[3]The ward is the smallest administrative subdivision in Nepal with an average area of $4km^2$ and an average population of 638.

[4]A GIS vector of known trails and footpaths can be obtained from the MENRIS mountain geoportal (http://geoportal.icimod.org/)

[5]The village development committee (VDC) is the second smallest administrative unit in Nepal and comprises an average of 9 wards.

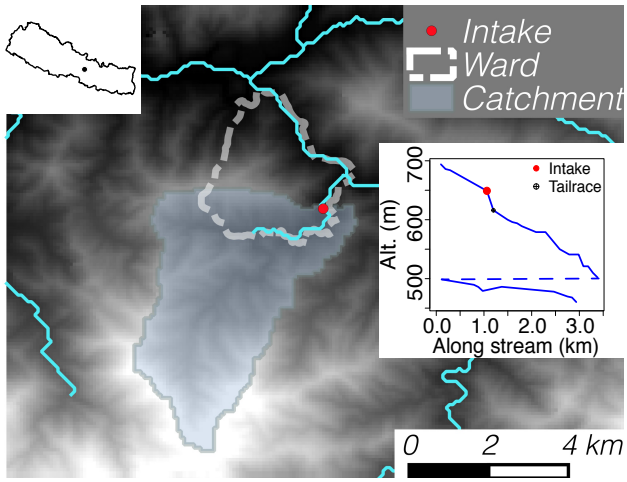


Figure 5.1: GIS algorithm for the supply shifter at Baireni ward 1 (Dhading district, Bagmati zone). (i)The probable location of streams (light blue) draining catchment areas larger than $3.5km^3$ were obtained from a high resolution digital elevation model through raster flow analysis (r.watershed in GRASS). (ii) Stream segments intersecting with the territory of each ward were clipped and subsetted. (iii) Altitude of regularly spaced points along the selected stream reaches were obtained using the digital elevation model. A separation of 30 meters between the point was selected as minimum penstock length (dark blue lines in inserted graph). (iv) Points with local maximum and minimum curvatures were identified based on the numerical derivatives of elevation. (v)The *minimum* curvature point associated (red dot) with the largest average slope to the nearest downstream *maximum* curvature point (black dot) was identified as the most appropriate location for micro-hydropower development. Once the location of interest identified (red dot), the area of the upstream topographic catchment (shaded in main map) was computed by topographic analysis (e.g., r.water.outlet in GRASS). Note that in the example the optimal intake is on a secondary stream with a much higher local slope than the main stream flowing north of the ward.

Note that P_{HH} is the present value at the construction of the infrastructure of all future regular (typically monthly) annuities paid by the households, which can be retrieved from P_{HH} by accounting for the interest rate and the use time of the infrastructure. This distinction does not affect my estimation of the price elasticity of demand, however, because the proportionality factor relating the present value and monthly annuities is absorbed in the intercept when regressing.

In addition to the REDB dataset, I also use secondary datasets derived from the Nepal Living Stantards Survey (NLSS), a household expenditure survey conducted in 2010 on a cross sectional sample of 3402 households in Nepal [166]. The first considered NLSS subset (GRID) consists of 82 grid-connected communities that were matched to the REDB dataset on population size (genetic matching [200]). All households from the considered subset are connected to the central electricity grid of Nepal and benefit from constant electricity tariffs.

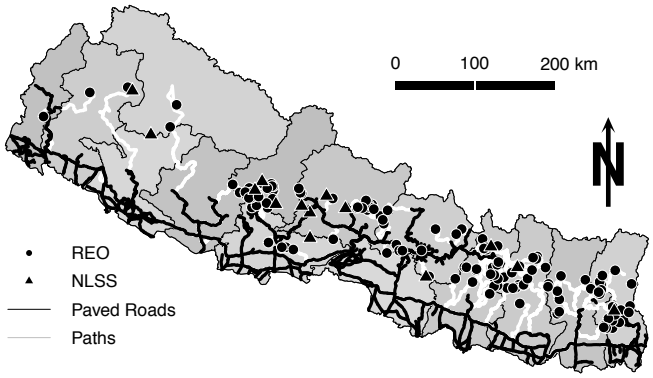


Figure 5.2: Approximate location and accessibility of the observed micro hydropower schemes.

The price paid by these households for electricity is uncorrelated to local micro-hydropower site conditions, which is a desirable identification property that will be exploited in the empirical strategy. Unit price and demand for electricity were derived from annual power expenditures and appliance ownership as described in Appendix D.1. The second NLSS subset, which consists of the 81 wards of the NLSS dataset that are not electrified, will be used in the discussion to check the robustness of the empirical strategy. Summary statistics of the three datasets are provided in Table 5.1 and the approximate location of the considered micro-hydropower schemes displayed in Figure 5.2.

5.2.2 Structural Equations

Consistent with the existing literature, I assume a double logarithmic demand curve for residential electricity (Equation (5.1)). Log-log functions are widely used to model household electricity demand [e.g., 83, 204, 24], mainly due to their empirical convenience: (i) they can readily be estimated through linear regressions, (ii) the estimated coefficients are then themselves estimates of the considered elasticities, that is the ratio of percentage changes in the demand and in the considered covariate, and (iii) the related standard errors provide measures of the variability of the estimated elasticities. Despite its widespread use, the double-log form does have a number of disadvantages. First, it imposes constant elasticities, while price elasticity has been observed to vary with income and the level of consumption [179]. This restriction is potentially problematic in the considered application and will be further addressed in the discussion. Second, log-log demand functions have ambiguous theoretical groundings [172], and can only be derived from consumer choice theory under very strong (and arguably often unrealistic) assumptions [68]. Nonetheless, the functional form of log-log demand can be consistent with utility-maximization theory in the specific case of unmetered electricity consumption by farm-households with no access to the labor market, as shown in Appendix D.2. Farm-household models consider utility-maximizing households, whose budget is constrained by income from their own production [19]. As such they are par-

	N	1.REDB (micro-hydro)	2.NLSS (Grid)	3.NLSS (Unconnected)
		$N = 121$	$N = 82$	$N = 81$
HH	284	45 138 220	101 170 262	70 110 136
R	283	18.31 45.23 79.09	0.98 2.79 7.09	11.16 22.20 96.01
kW	121	4 14 23		
C	121	772 2583 4266		
S	121	286 1007 1850		
P	203	84.35 102.25 135.44	0.63 0.86 1.20	
D	203	89 101 113	1212 1370 1712	
I	156		23 49 75	17 32 60
pBiz	284	0.067 0.096 0.172	0.167 0.333 0.417	0.000 0.000 0.014
houseVal	147		1.9 2.9 4.2	1.1 1.5 2.2
wage	144		11 15 20	10 12 16
A	203	22 31 164	22 28 104	
Precip	121	2392 2653 3206		
Yr	121			
2008		19% (23)		
2009		21% (25)		
2010		35% (42)		
2011		26% (31)		

a b c represent the lower quartile a, the median b, and the upper quartile c for continuous variables.
N is the number of non–missing values.
Numbers after percents are frequencies.

Table 5.1: Summary statistics of the sampled wards by dataset/electricity source. HH: Number of electrified household, R: Distance to nearest road head (km), kW: Capacity of the micro-hydropower scheme (kW), C: Total construction costs (thousands of 2005 Nepalese Rupees), S: Total construction subsidies (thousands of 2005 NRp), P: Subsidized electricity price (thousands of NRp per kW per household), D: Electricity consumption (W per household), I: Annual income (thousands of 2005 NRp), pBiz: ratio of households owning a non agricultural business, houseVal: Average value of dwelling (thousands of 2005 NRp), wage: average day laborer's wage (tens of 2005 NRp per day), A: Upstream catchment area at the most promising micro-hydropower site, Yr: Construction year. N designate communities for columns 1 (REDB-micro-hydro) and 2 (NLSS-Grid) and households for column 3 (NLSS-Unconnected)

ticularly appropriate in developing countries, where a substantial share of households (96.8% in Nepal [166]) earn at least part of their livelihood through work in their own enterprise. In that context, electricity consumption affects at once the utility of the household and the income of the farm and the derived demand incorporates both the residential and productive uses of electricity.

On the supply side, hydropower is characterized by significant sunk costs: initial investments typically dominate operation expenses [4], which are here neglected. Following Aggidis et al. [4] and others [11, 205, 206], I model the cost function of micro-hydro as a power law of the capacity of the plant and a series of covariates related to infrastructure efficiency and

local site conditions. Micro-hydro revenue are proportional to the capacity of the plant with the market price per unit of capacity as a proportionality constant (recall that in the absence of meters, households are billed on a capacity basis). Although their bulkiness and capital intensity create optimal conditions for a natural monopoly, local utilities are typically subject to public pricing [16]. Regulating prices is necessary for utilities to achieve economic efficiency and equitable access, especially for the supply of electricity, which turns out to have a strongly price-inelastic demand (Table 5.3). I here assume average cost pricing because all the plants in the considered sample are community owned. Average cost pricing allows full cost recovery, while keeping the price low enough to prevent the infrastructure from being underutilized [16]. Micro-hydro costs are exactly compensated by revenues, which is consistent with the cost recovery assumption used to construct the dataset in Section 5.2.1. Under these conditions, the average total cost function of the micro-hydro can also be represented by a log-log function.

Considering the above arguments and the available information, the following system of simultaneous structural equations is used in the estimation:

$$\begin{aligned}\ln kW_{HH} &= \gamma_0 + \gamma_P \ln P_{HH} + \gamma_R \ln R + \gamma_{HH} \ln HH + \gamma_{pr} \ln precip \\ &\quad + D_{zon,D}\delta_{zon} + D_{Y,D}\delta_Y + \gamma_{pBiz} \ln pBiz + \varepsilon_D \qquad (5.1)\\ \ln P_s &= \alpha_0 + \alpha_{kW_{HH}} \ln kW_{HH} + \alpha_R \ln R + \alpha_{HH} \ln HH + \alpha_{pr} \ln precip \\ &\quad + D_{zon,S}\delta_{zon} + D_{Y,S}\delta_Y + \alpha_A \ln A + \varepsilon_S \qquad (5.2)\end{aligned}$$

where ε_D and ε_S are idiosyncratic random errors and α_i and γ_i (with $i \neq 0$) are elasticities of the supply and demand respectively. Of particular interest are the price elasticities $\alpha_{kW_{HH}}$ and γ_P, which are the respective slopes of the supply and demand curves represented in a double logarithmic scale, and the scale factor $\alpha_{kW} = \alpha_{kW_{HH}} + 1$, which represents the ratio of percentage changes in the cost and size of the infrastructure (i.e. a small α_{kW} represents large economies of scale).

The utility maximization model (Appendix D.2) suggests that, in addition to price, electricity consumption is driven by the income determinants of the farm household, that is their endowment in production capital and their ability to generate profits from the consumed electricity. Evidence found in the literature further suggest that demand for electricity is affected by characteristics of households (size [83], appliance ownership[239, 145], type of dwelling[83], etc), individuals (e.g., eduction [168], age[83]) and communities (e.g., size[83]). I consider the following covariates in the demand specification (Equation (5.1)):

- R: Remoteness in km to the nearest road, which affects the households' access to education and markets – both to generate income and buy electrical appliances.
- HH: Community size is expected to affect electricity consumption through its effect on electricity quality and access to electrical appliances [83]. In Nepal, community size is also a key determinant of electrification subsidies (see Appendix D.3), which influence the electricity price actually faced by the households.

- *precip*: Precipitation is an obvious determinant of income in agricultural communities.
- $pBiz$: Denotes the propensity of households owning non-agricultural businesses, which as argued in Section 5.2.3 is strongly correlated to their electricity demand.

Covariates of the supply equation (Equation (5.2)) relate to site conditions liable to affect the cost of micro-hydropower infrastructure. Such conditions include (i) head and flow conditions, which affect the type and efficiency of the turbine, (ii) local topography through its effect on the cost of civil works, (iii) site accessibility that determines transportation costs and (iv) local material and labor costs. These aspects are accounted for in the supply specification as follows:

- R: Remoteness is taken as a proxy for site accessibility.
- HH: Community size affects labor costs through the provision of local labor – anecdotal evidence point toward a decreases of total construction costs by 12 to 25 percentage points [180].
- *precip* and A: Precipitation and the area of the contributing watershed both affect the flow volume available for hydropower production.

Finally, regional variations in price and consumption and temporal trends are captured by dummy variables (δ_{zon} and δ_Y) denoting the construction year of the infrastructure and the administrative zone of the community[6].

5.2.3 Identification

It is immediately apparent from Equations (5.1) and (5.2) that price is simultaneously determined by supply and demand considerations. Under these conditions, using a standard ordinary least squares (OLS) estimation yields substantially biased electricity demand elasticities, as shown in Dubin [74]. Although these endogeneity concerns can be resolved using instrumental variables, finding good instruments is a significant challenge because it requires identifying 'clean' and significant sources of variation that predict each individual endogenous variable, while not being correlated to the others. In my system this requires variables that are strongly correlated to either the unit cost of micro-hydro or household electricity demand, while being uncorrelated to the other. The considered dataset allows constructing two such variables: A and $pBiz$. Arguments supporting their relevance (i.e. the *first stage*), that is their respective correlation to electricity cost and household electricity consumption, are provided in Section 5.2.2 above. Their validity (i.e. the *exclusion restriction*) can be argued as follows:

- A is a topographic characteristic that describes the hydropower potential at *one particular site* within each ward's territory. This site is determined by the local slope and

[6]There are 14 administrative zones in Nepal extending over an average of 2600 wards each

curvature of the river bed and is likely uncorrelated to any determinant of electricity demand. If at all affected by topography (e.g., through agricultural productivity or accessibility), demand is more likely related to the variables describing *average* topographic characteristics of the community's territory.

- *pBiz* was recorded in 2001 Census of Nepal, *prior* to the construction of the considered micro-hydro infrastructure, which were all built between 2008 and 2011. The probability of owning a business then is not affected by the price of the (still inexistent) micro-hydro scheme.

Residual threats to both exclusion restrictions are addressed in Section 5.4.3 through robustness checks using the NLSS dataset.

5.2.4 Estimation

Conditional on a strong first stage and a valid exclusion restriction, a two stage least square (2SLS) procedure produces consistent estimators for the elasticities [240]. Yet 2SLS is inefficient in the considered application, where the disturbances of the simultaneous structural equations are contemporaneously correlated. In that context, a feasible generalized least squares version of the 2SLS estimation leads to asymptotically more efficient estimates. Unfortunately, unlike 2SLS, the standard three-stage least squares estimator (3SLS) described in Zellner and Theil [250] is only consistent if all disturbance terms and all instrumental variables are uncorrelated [197]. In this paper I use the GMM-3SLS estimator defined by Amemiya [10], which has been shown to be both consistent and asymptotically efficient in the estimation of system of simultaneous equations with different sets of instruments for each endogenous variables [197]. The GMM-3SLS estimator is provided in Appendix D.4 with the underlying formal assumptions. Just like 2SLS, GMM-3SLS is at risk of small sample bias, especially in the presence of multiple weak instruments. However, 2SLS is approximately unbiased for just identified instruments if the first stage is not too weak ($F \geq 5$ [115] or $t > 1.5$ [15]) and if the exclusion restriction holds.

5.3 Results

Ordinary Least Squares (OLS) regressions allow assessing the significance of control variables and their effect on the correlation between price and capacity – the relation of interest in this study. As expected, the demand specifications (Table 5.2: (1-3)) show a negative and significant correlation between the unit price faced by households and their level of consumption. The correlation remains significant when controlling for all considered covariates and its magnitude increases. The largest single increase in magnitude occurs when controlling for remoteness, which is itself negatively and significantly correlated to consumption. This is consistent with my presumption that remote communities consume less and face higher unit costs. On the supply side (Table 5.2: (4-6)), specification (5) shows that unit costs decrease

with infrastructure size but increase with community size. Yet because consumption is correlated to community size, larger infrastructure are built in larger communities so the two correlations compensate almost exactly in Table 5.2 (4). The negative correlation between infrastructure size and unit costs is expected due to economies of scale and persists when controlling for all considered covariates (6). The positive correlation between community size and unit price is somewhat surprising but loses its statistical significance when accounting for the simultaneity bias (Table 5.3: (4)).

	Dependent variable:					
	log(kW.HH)			log(Ps)		
	[Demand]			*[Supply]*		
	(1)	(2)	(3)	(4)	(5)	(6)
log(P.HH)	-0.092^{***}	-0.097^{***}	-0.082^{**}			
	(0.032)	(0.031)	(0.038)			
log(kW)				0.003	-0.477^{***}	-0.299^{**}
				(0.031)	(0.147)	(0.138)
log(R)		-0.030^{*}	-0.037^{*}	0.006	-0.023	0.009
		(0.015)	(0.021)	(0.029)	(0.029)	(0.031)
log(HH)			0.040^{*}		0.519^{***}	0.296^{**}
			(0.024)		(0.156)	(0.148)
log(Precip)			-0.067			0.030
			(0.063)			(0.095)
log(pBiz)			0.060^{**}			
			(0.024)			
log(A)						0.023^{**}
						(0.009)
Observations	121	121	121	121	121	121
Adjusted R^2	0.058	0.079	0.264	-0.016	0.063	0.401
Note:						*p<0.1; **p<0.05; ***p<0.01
						Standard Errors are in parentheses
						Intercept Included in all specifications
						Yr and ZONE fixed effects included in (3) and (6)

Table 5.2: Ordinary Least Squares Specifications for Demand (1-3) and Supply (4-6)

Crucially, the two instruments are strongly correlated to their respective endogenous variable. As expected, the propensity for owning a business ($pBiz$) is positively correlated to household electricity consumption. The positive correlation between the area of the contributing catchment (A) and the unit cost of electricity is consistent with the fact that low-head (and therefore high-flow) hydropower is comparably more costly for a given amount of power [169]. The two instruments are relevant, with t-values of 2.7 (A) and 2.5 ($pBiz$) and partial F statistics of 7.2 (A) and 6.2 ($pBiz$) in the first stage specifications (Table 5.3 (1-2)).

	First Stage		3SLS	
	log(P.HH) Demand	log(kW) Supply	log(kW.HH) Demand	log(Ps) Supply
	(1)	(2)	(3)	(4)
log(P.HH)			−0.1394	
			(0.1503)	
log(kW)				-0.0659
				(0.58352)
log(A)	0.043***	−0.006		0.0244**
	(0.016)	(0.006)		(0.00981)
log(pBiz)	−0.006	0.061**	0.0598**	
	(0.060)	(0.024)	(0.0242)	
log(R)	0.0003	−0.036*	−0.0360*	0.0184
	(0.053)	(0.021)	(0.0213)	(0.58352)
log(HH)	−1.059***	1.044***	−0.1037	0.0515
	(0.059)	(0.024)	(0.1641)	(0.61285)
log(Precip)	0.146	−0.085	-0.065	0.045
	(0.165)	(0.067)	(0.064)	(0.103)
Observations	121	121	121	121
Adjusted R^2	0.869	0.974		

Note: *p<0.1; **p<0.05; ***p<0.01
Standard Errors are in parentheses
Intercept and Yr and ZONE fixed effects included in all specifications

Table 5.3: First Stage and Three Stage Least Squares Estimations. Instruments A and $pBiz$ have partial F stats of 7.2 and 6.2 respectively on the first stage. The scale factor for micro-hydropower costs can be obtained by adding 1 to the price-elasticity of supply in specification (4): $\alpha_{kW} = 0.93$ (t-score=1.61, p-value=0.11)

The final 3SLS specifications (Table 5.3: (3-4)) use the instruments to control for simultaneity biases and identify the respective causal effects of price on consumption (3), and of infrastructure capacity on unit costs (4). The estimated price elasticities are -0.14 (90%CI: $[-0.39, 0.11]$) for demand and -0.07 (90%CI: $[-1.03, 0.89]$) for supply. Also noteworthy are the small positive effects of remoteness and community size on infrastructure price (4), and the negative effect of remoteness on electricity consumption (3). With the exception of that last coefficient, none of the estimated elasticities are statistically significant at the 90% level – a likely consequence of the relative weakness of the first stage [15]. Despite their weak statistical significance, the obtained values can nonetheless be regarded as reliable estimates of the considered elasticities, as argued in the discussion below.

5.4 Discussion

5.4.1 Supply

The 3SLS supply specification (Table 5.3 (4)) reveals no significant effect of infrastructure size on its unit costs, which denotes little economies of scale. The estimated scale factor[7] of 0.93 is larger than the typical values (0.60-0.82) compiled in Aggidis et al. [4] for US and European schemes. This discrepancy in Nepal can be explained by the provision of local materials and community labor, which externalizes a significant part of the fixed costs related to civil works (excavation, structures). This translates into lower economies of scale with respect to the capacity of the plant and a higher scale factor. I substantiate this claim through case studies with detailed cost estimates of the four Nepalese micro-hydropower projects presented in Figure 5.3. Scale factors are obtained by computing the ratio of variable costs to total costs (before taxes) and yield a median value of 0.90, which is comparable to the 3SLS-estimated value.

The coefficient for remoteness reported in Table 5.3 (4) is noisy and of low amplitude, which is somewhat surprising because one might expect the extreme landscape to have a significant effect on transportation costs in Nepal[8]. The OLS regressions presented in Table 5.4 show that remoteness ceases to significantly affect infrastructure price when controlling for the size either the community or the infrastructure, suggesting that micro-hydropower is only feasible in remote locations when infrastructure size (and therefore community size) is large enough. Absent significant economies of scale that would explain that pattern, this effect is likely associated to a subsidy structure that favors large infrastructure in remote areas (see Appendix D.3). Two further reasons can be mentioned for the low amplitude of the remoteness coefficient. First, I measure remoteness as the distance along footpaths to the

[7] A scale factor of 0.93 means that doubling the capacity of the infrastructure increases its total cost by 93%.

[8] For instance, Jacoby [120] found a significant causal effect of accessibility on local wages and land prices; Khennas et al. [128] reports anecdotal evidence of cases, where that transport covers as much as 25% of total micro-hydropower costs

nearest drivable road, without accounting for the topography. Yet in mountainous terrain slope affects speed, transportation times and ultimately transportation costs. Secondly (and perhaps more relevantly), a large part of the transportation costs may be internalized by communities. The REDB dataset reports the total external costs and it can very well be that a significant portion of the transportation costs was provided in nature by porters from the communities. Available case studies [180] typically allocate about 50% of total transportation costs to labor contributions from the community. Resulting observed elasticities for remoteness (median: 0.04) exclude local labor costs (e.g., porters) and are on the order of the 3SLS estimated value of 0.02.

		Buwa	Nurkhuwa	Suntale	Yari	
Salient Features						
Connected Households	HH	274	312	1094	230	
Capacity	kW	18	31	100	23	
Design Flow	lps	84	140	215	122	
Head	m	39	41	87	35	
Remoteness	km	40	40	80	20	
Turbine Type		Cross Flow	Cross Flow	Pelton	Cross Flow	
Tariff						
Domestic	NRp/W/mth	3	2	3	2	
Industrial	NRp/kWh	10	8	12	10	
Subsidies	kNRp	2,475	3,875	15,500	2,990	
Costs						
Monetary (i.e. non local) costs	kNRp	5,622	7,183	32,487	6,170	
Civil	%	7.7	8.6	6.5	12.8	
Electro Mech.	%	64.5	65.8	68.0	65.7	
Transportation	%	4.6	4.0	10.6	1.9	
Taxes and services	%	23.2	21.7	14.9	19.5	
Scale factors						**MEDIAN**
kW*	-	0.900	0.891	0.923	0.841	**0.895**
R**	-	0.046	0.040	0.106	0.019	**0.043**
HH***	-	0.104	0.142	0.243	NA	**0.142**

* Ratio of variable (EM and Tranp.) against total construction costs before taxes (EM, transp and civil)
** Ratio of transportation against total costs before taxes
*** Ratio connection and transmission costs against total costs before taxes.
Source: AEPC feasibility studies (2009)

Figure 5.3: Detailed cost breakdown of four micro-hydropower projects in Nepal [180].

	Dependent variable:					
	log(C)	log(C/HH)	log(C/kW)	log(kW/HH)	log(kW)	log(HH)
	(1)	(2)	(3)	(4)	(5)	(6)
log(R)	0.397***	−0.024	0.002	−0.027	0.395***	0.422***
	(0.148)	(0.031)	(0.032)	(0.026)	(0.136)	(0.126)

Note: *p<0.1; **p<0.05; ***p<0.01
Construction year fixed effects included

Table 5.4: Remoteness elasticity of several infrastructure characteristics: (1) construction costs, (2) costs per household, (3) costs per kW (P_s), (4) capacity per household ($kW_H H$), (5) capacity, (6) community size.

5.4.2 Demand

The price elasticity of demand obtained in this study ($\alpha_P = -0.14$) is at the lower end (in absolute value) of the range of elasticities estimated in past studies, as seen on Figure 5.4. In particular, my estimate is lower previously estimated values in India based on macro- ($\alpha_P = -0.63$) [35] and micro-data ($\alpha_P = -0.29$ during the dry season) [83]. Part of the discrepancy probably relates to important differences in the attributes of the consumed electricity: connections are here unmetered (i.e. the elasticity concerns peak demand) and mini-grid electricity is (arguably) of lesser quality than grid electricity in India. Yet an important characteristic of my dataset is the particularly low level of consumption of the sampled communities, with a median household capacity of 101W against approximately 1370W for comparable grid connected households (Table 5.1). This feature sets this analysis apart from previous studies and likely explains the lower estimated elasticity.

Indeed, a key disadvantage of log-log demand function is their reliance on the assumption of constant elasticity, which has little theoretical grounding for electricity [172] . While the farm-household model derived in Appendix D.2 does predict a constant price elasticity of demand, the predicted elasticity is larger (in absolute value) than one[9], which does not reflect typical elasticity ranges found in empirical studies (Figure 5.4). This mismatch is caused by the Cobb-Douglass (power law) function used to model farm production and household utility. Such a functional form is necessary to obtain a log-log demand function but assumes a unit elasticity of substitution between the production factors. In other words, electricity is neither a substitute nor a complement to the other production factors, which is, of course, unrealistic[10]. While other production and utility functions can be used to relax the unit substitution elasticity assumption (e.g., the constant elasticity of substitution function used in Agostini and Saavedra [5]), the resulting demand function, when at all analytically

[9]Positive returns to scale in the production function requires $\mu \geq 0$ and therefore $\gamma_P = \frac{1}{\mu - 1} \notin]-1, 0[$

[10]Electricity is (to some extent) a substitute to labor and a complement to capital.

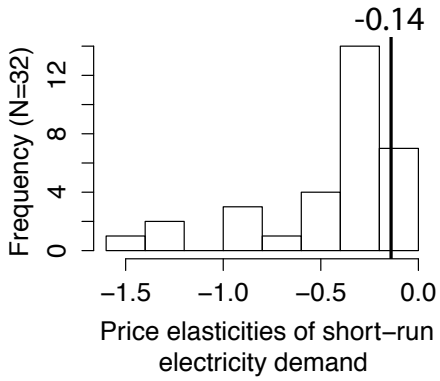


Figure 5.4: Histogram of short-run price elasticities of electricity demand estimated in past studies. The considered estimates are from a 32 studies covering 8 countries (USA, UK, Switzerland, Australia, India, Paraguay, Isreael and Canada) and were compiled by Bendezú and Gallardo [26] and Fan and Hyndman [81]. With a value of -0.14, this study's own estimate for mini-grid electricity in rural Nepal is at the lower and (in absolute value) of the spectrum.

tractable, entails considerable estimation challenges when taken to empirical data.

These theoretical shortcomings were confirmed empirically [e.g. 179]. In particular, elasticity is liable to decrease (in absolute value) at extreme levels of consumptions. At the lower end of the spectrum, price has generally little effect on households' minimum level of consumption [198] and electricity effectively becomes close to a vital good with no near substitute[11]. This effect is the most likely explanation to the low elasticity estimated here for Nepal and is clearly evidenced in the available dataset, as shown in Figure 5.5, where the estimated elasticity increases when discarding the lowest-consuming households. Discarding the 12 communities (10% of the sample) with the lowest levels of consumption raises demand elasticity to -0.22 and makes it nearly significant at the 90% level with a confidence interval of $[-0.43; 0.03]$. These results reveal an interesting tradeoff to be made when using double-log functions to estimate elasticities for extreme (low or high) levels of electricity consumptions: while sufficient consumption variability is necessary in the sample to increase the accuracy of the estimate, too much variability may violate the constant elasticity assumption attached to the log-log formulation. In that context, the proposed resampling analysis (Figure 5.5) can be used as a robustness check to assess the relative importance of these two effects.

[11]For instance, a very small and non-substitutable amount of electricity is needed to recharge batteries for portable torch lights

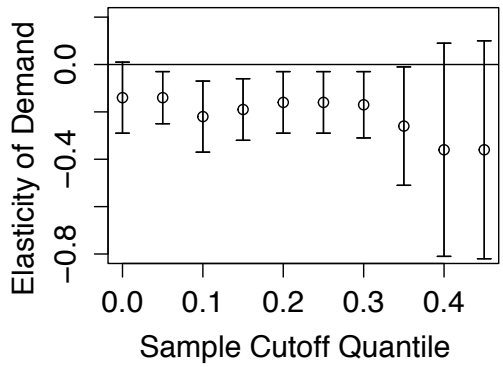


Figure 5.5: Estimated (3SLS) price elasticity of electricity demand on sub-samples of the REDB dataset (standard errors given as error bars). Subsamples are drawn by selecting communities with an average level of consumption higher then the chosen sample cutoff quantiles. The graph shows that demand elasticity increases as lower-consuming households are discarded, which evidences the effect of consumption on demand elasticity.

5.4.3 Instruments

The first stage coefficients of the considered instrumental variables (Table 5.3 (1-2)) are of the expected sign (see Section 5.3), but fairly weak with partial F stats of 6.2 ($pBiz$) and 7.2 (A), and therefore potentially subject to weak instrument biases [41]. Three arguments can nonetheless be made to support the relevance of the instruments. First, the system is just-identified, which causes 3SLS to be *approximatively median unbiased* for relatively weak instruments [15]. Second, when used in a 3SLS estimation, the resulting elasticities reproduce well previous findings from the literature and detailed case studies in Nepal, as discussed in Sections 5.4.1 and 5.4.2. Finally, because the weak instrument bias is in essence a small-sample issue, I test the robustness of my estimates to decreasing sample sizes by applying the estimation procedure on random subsamples of the original dataset. As seen on Figure 5.6, the resulting mean 3SLS estimates do not change signs and stay within a fraction of a standard error distance from the full sample 3SLS estimates. This suggests that the estimated elasticities are not subject to small-sample biases.

Unlike first stage correlations, exclusion restrictions cannot be directly tested in a quantitative manner. For the demand instrument (A), I build confidence in the validity of the exclusion restriction by using a matched NLSS dataset of grid-connected households to show that in the absence of micro-hydropower infrastructure, the instrument is not significantly related to electricity demand or expenditure (Table 5.5). Assuming that the matched NLSS sample is representative of the original dataset, this shows that local site conditions only affect electricity demand through their effect on infrastructure costs. For the supply instrument ($pBiz$), the exclusion restriction rests on the time difference between the observation of the instrument (2001) and the construction of the hydropower infrastructure (between 2008 and 2011). This rules out reverse causality because the propensity of owning a business in

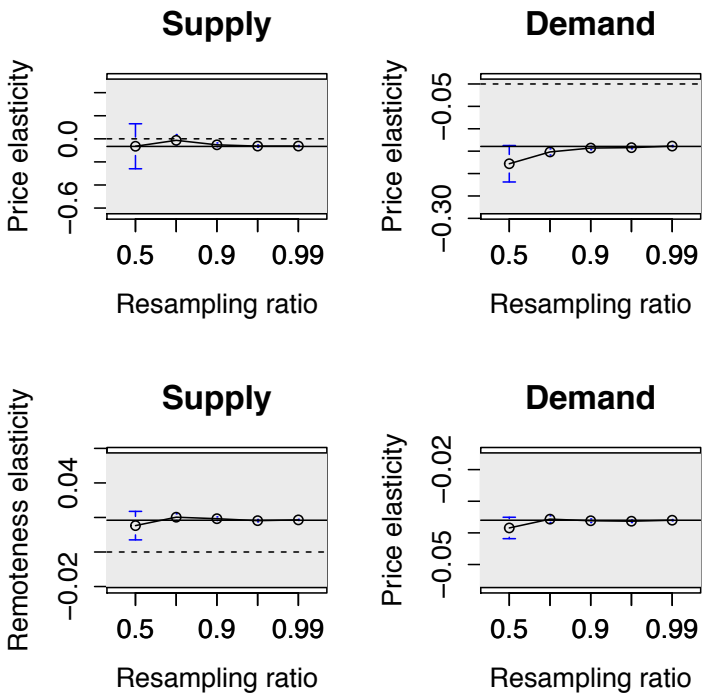


Figure 5.6: Sensibility of 3SLS estimates to sample size. 3SLS estimates of the price (row 1) and remoteness (row 2) elasticities of supply (column 1) and demand (column 2), against the size of the subsample (expressed as a ratio of the original sample size). Subsamples are drawn randomly (no replacement) and the mean 3SLS estimates (dots) and the standard errors over all 1000 resampling repetitions are displayed in the graphs. Full sample estimates (horizontal line) and their standard errors (grey band) are represented on the graphs as well. In all considered cases, the 3SLS estimates from the random subsamples do not change signs and stay within a fraction of a standard error distance from the full sample estimates.

2001 is certainly not affected by the price of the (still inexistent) infrastructure. However, it is still possible for an underlying cause to affect both the instrument and the future cost of the infrastructure. For instance, average rainfall affects hydropower costs, but may also affect households' likelihood to engage in non agricultural activities. I minimize that risk by controlling for $precip$, HH, R and regional fixed effects in the 3SLS estimation. Finally, $pBiz$ may also have a direct effect on infrastructure costs through its effect on the local economy. In particular, the ratio of non-agricultural households may be correlated to local wages or construction costs, which in turn will affect the cost of the infrastructure. Although the available dataset does not allow to control for that possibility, I build confidence in the validity of the exclusion restriction by using the NLSS dataset to regress the ratio of non-agricultural households against average dwelling prices and daily laborers' wages. In order to emulate communities' situation before the construction of the hydropower plant, I only consider the

subset of NLSS wards that do not have access to to grid or micro-hydropower electricity. Results in Table 5.6 show no significant correlation, which supports the hypothesized exclusion restriction.

	Dependent variable:	
	log(kW.HH)	log(P.HH)
	OLS	*OLS*
	(1)	(2)
log(R)	-0.063^{***}	-0.112^{**}
	(0.014)	(0.046)
log(HH)	0.094^{***}	-0.003
	(0.035)	(0.115)
log(A)	-0.007	0.015
	(0.007)	(0.022)
log(pBiz)	0.061^{**}	0.013
	(0.025)	(0.080)
Observations	82	82
Adjusted R^2	0.467	0.065
Note:	*p<0.1; **p<0.05; ***p<0.01	
	Intercept included	
	Yr and ZONE fixed effects included	

Table 5.5: Exclusion Restriction for A. First Stage Reduced form regressions for the demand on grid connected community. In the absence of micro-hydropower infrastructure the supply shifter A does not have a significant effect on household demand or unit price.

	Dependent variable:	
	pBiz	
	OLS	
	(1)	(2)
houseVal	-0.022	
	(0.020)	
wage		-0.004
		(0.004)
Observations	70	62
Adjusted R^2	0.003	0.005
Note:	*p<0.1; **p<0.05; ***p<0.01	

Table 5.6: Exclusion Restriction for $pBiz$. There are no significant correlation between the ratio of non agricultural households and house prices or daily laborers' wages.

5.5 Conclusion

This chapter presents an econometrically robust approach to identify and estimate key economic design inputs for off-grid electrification in mountainous areas: the price elasticity of electricity demand and the scale factor of micro-hydropower generation. The approach, which uses commonly available salient features of existing infrastructure and instruments derived from remote sensing (topography) and census data is particularly applicable to developing countries, where detailed consumption data are missing and the absence of meters on household connections prevalent. Despite large standard errors, the resulting estimates of -0.14 (price elasticity of demand) and 0.93 (scale factor of micro-hydropower) are in line with detailed case studies in Nepal. My findings assert community labor contributions and low levels of consumption, two fundamental aspects of off-grid electrification in developing countries, as having a notable effect on the estimated elasticities that separates them from previous findings. This study underlines the need to keep these particularities in mind, when using empirically estimated price elasticities to design and managed distributed mini grids in developing countries.

Chapter 6

Micro Hydro [em]Power: A WebGIS tool for community-based design and large-scale feasibility mapping.

6.1 Introduction

Access to electricity remains an impediment to development in many parts of the world. According to the IEA [112], more than 1.3 billion people lack access to electricity. 85% of these people live in rural areas with low population densities and prohibitive grid extension costs. In that context, decentralized distributed generation (DDG), whereby electricity is generated at the point of consumption, stands out as a promising and affordable strategy for rural electrification [159]. Community-scale run-of-river hydropower – *micro-hydropower* – is a particularly attractive DDG option in mountainous regions, where appropriate slope and runoff conditions are encountered.

In a typical micro hydropower setup (Figure 6.1), water is diverted from a stream through an intake at a small weir, a man-made barrier across the river that maintains a continuous flow through the intake. The diverted water flows along a free surface headrace canal into a small storage structure called the forebay, where it slows down sufficiently for suspended particles to settle. The forebay feeds into the penstock, a pressure pipe conveying the water to the powerhouse, where one or more turbines convert the kinetic energy of the water into electricity. To minimize energy losses to friction and maximize the kinetic energy of water entering the turbine, the infrastructure is laid out so as to minimize the slope of the canal and maximize the slope of the penstock. Consequently, local topography has a strong influence on the ideal placement of infrastructure components, though other local conditions such as accessibility and proximity to other existing infrastructure have a significant effect as well. Thanks to the low level of technology of its components, micro hydropower often emerges as the most cost effective DDG option for mountain communities [169]. Unlike conventional hydropower, micro-hydropower has a limited impact on the landscape and on the flow regime

of the stream: it does not store significant volumes of water and releases diverted waters at a short distance from the intake.

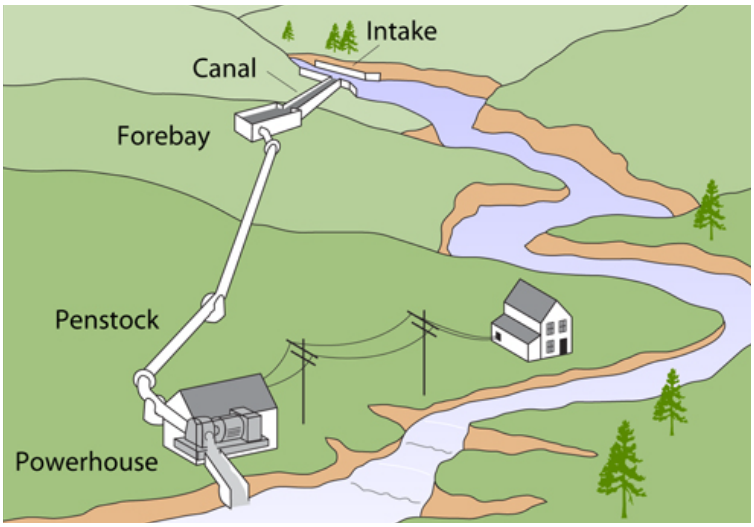


Figure 6.1: Typical setup of a micro hydropower plant. Source: Office of Energy Efficiency and Renewable Energy, United States Department of Energy.

Despite their promise, micro hydropower programs have had mixed success. For example, even in Nepal, a country with huge hydropower potential, low rural electrification rates, favorable policies and local hydropower expertise, the technology has difficulties scaling. After 50 years of implementation, micro hydro currently supplies less than 4% of off-grid households and up to 30% of existing plants are not in operation [128]. Factors including poor design, inappropriate management and a low level of local participation have been brought up to explain these poor outcomes [e.g., 128, 151]. Yet fundamentally, as elaborated below, these factors emerge from a major information gap between the urban centers, where funding agencies and technical expertise are concentrated, and the rural communities that will benefit from micro hydro implementation.

First, there is an acute shortage of data about existing micro hydro infrastructure to inform policy [73]. Because micro-hydropower plants typically fall below the minimum plant capacity requiring a license [87], the number of systems that have been built and are still functioning in unknown. Without implementation data, policy makers have to rely on computer tools identify promising regions for new micro hydro installations. Although a number of large-scale Geographic Information Systems (GIS) are available to map hydropower potential using stream and topography information [see review in 178], these tools typically evaluate infrastructure feasibility based on economic assumptions appropriate to grid-connected infrastructure, where the generated electricity is sold to power utilities at fixed feed-in tariffs, and are hardly applicable to off-grid electrification.

Second, local streamflow data are missing to properly design infrastructure. The design of micro-hydropower plants relies on optimizing the plant capacity based on knowledge of

streamflow variability, as described by the flow duration curve of the stream. Flow duration curves are difficult to predict *a priori*, particularly in hydrologically complex Himalayan catchments. The underlying challenge of predicting flow duration curves in the absence of measured local streamflow is a component of the global "Prediction in Ungauged Basins (PUB)" movement and represents a major contemporary challenge in hydrologic research [32]. Promising approaches for Nepal are described in Chapters 2 to 4 of this dissertation.

Third, knowledge of the electricity demand of local communities is essential to establishing the economic feasibility of micro hydro infrastructure. Electricity generated off-grid by micro hydro plants cannot be exported to urban centers, meaning that the feasibility is strongly coupled to local demand. Significant recent research has been devoted to the estimation of the demand curve for residential electricity [see 80, 106], but the absence of metered consumption data for off-grid electricity in rural Nepal prohibits the use of standard demand estimation techniques. Chapter 5 outlines a method to characterize electricity demand using publicly available infrastructure surveys. The high level of uncertainty associated with the estimated (mean) elasticities suggests that the available data does not allow accurate predictions of electricity demand curves at the local level.

Finally, technical knowledge is challenging to access for would-be infrastructure promoters in rural communities [85, 87], whose active involvement in the planning process is widely recognized as a key prerequisite to success [128]. Although several decision support systems have been developed to help off-grid communities evaluate the technical and financial viability of potential clean energy projects [e.g., 31, 142], their use requires technical skills and computational equipments that are hardly available in rural Nepal.

The increasing availability of information and communication technologies (ICT) offers a pathway to overcome informational barriers: 60% of households in rural Nepal owned a mobile phone in 2011 [54]. Recent developments in open-source, web-based and participatory GIS [215] can be leveraged to both access local knowledge and disseminate expert technical advice to remote areas. To date, the most promising applications have been for health-related information systems [134, 150]. These technologies have yet to be used to facilitate rural electrification. Micro Hydro [em] Power, the tool presented in this chapter, addresses the four key information barriers affecting micro-hydropower development. First, it uses global remote sensing data to map the feasibility of micro hydropower for off-grid electrification at the national scale. Second, the tool uses the most recent hydrological modelsChapters 2 to 4 to estimate streamflow durations at the local level for micro hydropower design. Third, the application assesses and records local costs, electricity demand and loan conditions, which are provided by the communities *via* an interactive web interface. Lastly, thanks to its open source and web-based nature, the tool can be accessible to local decision makers in rural communities, providing them with the necessary information to site, design and evaluate the feasibility of micro-hydropower with no technical background.

Although the examples here are drawn specifically from the experience in Nepal, similar opportunities and challenges relating to micro hydro arise worldwide, Techniques to address information barriers and facilitate micro hydro implementation could thus, ultimately, reach a global audience. This research focuses on developing a tool for use in Nepal with the

expectation that the methods would provide a prototype for other similar regions.

The first part of the chapter provides a detailed description of the tool, outlining its software components (Section 6.2.1), graphical user interface (Section 6.2.2) and key algorithms (Section 6.2.3). The second part of the chapter (Section 6.3) evaluates the tool's performance by assessing its ability to predict the location and characteristics of a sample of 148 existing micro hydro schemes in Nepal.

6.2 Tool Description

6.2.1 Software components and data

All software components are open source, with licenses providing users with the freedom to run, modify and freely redistribute the original or modified program without further limitation or royalty payments [138]. Open source software is widely recognized for its ability to close the digital divide between rich and poor countries by increasing access and encouraging local developments [157, 158, 196]. To ensure software components could interface and work together, the R software environment for statistical computing was selected as the main computational engine. R has strong GIS capabilities with seamless interfaces to the Geospatial Data Abstraction Library (GDAL) [127], the Geometry Engine Open Source (GEOS) [30] and the Geographic Resources Analysis Support System (GRASS) [29, 219], which respectively enable it to manage spatial vector data, perform geospatial vector operations and store and analyze raster data. Recent developments allow interactive web applications to be built with little to no prior HTML knowledge, and open-source interactive maps[1] can be integrated in the web-based user interface using the Leaflet JavaScript library [3].

The tool relies on publicly available data (Table 6.1), which are stored internally as RData files for vector objects, or as sqlite tables within GRASS for raster objects.

6.2.2 GUI

The Graphical User Interface is accessible on a web browser (`https://mfmul.shinyapps.io/mhpower/`) and consists of an interactive map and a tab panel that can be toggled between two display modes. The *Local Designer* mode is intended to be used by local communities to assess the feasibility of micro-hydropower within their locality. The community (i.e. the VDC[2]) of interest can be selected on the interactive map, and the five most topographically suitable micro hydro layouts in close proximity to that community are determined (Figure 6.2 (b)). A slider panel allows cost, demand and financial parameters to be adjusted to reflect local conditions (Figure 6.2 (c)). Locally tuned parameters are subsequently reviewed and included into the tool if the user chooses to submit them. The interactive map allows users to select individual micro hydro layouts and to display the salient features, costs and

[1] Free interactive background maps are available on `https://www.mapbox.com` (Accessed in April 2015)

[2] VDCs are the second lowest administrative subdivision in Nepal, roughly equivalent to US counties

Name	Description	Source and Units
GDEM ASTER v2	Digital elevation model Resolution: 1' (30m) Coverage: global	NASA, METI [225]
TRMM 3B42 v7	Daily gridded precipitations Resolution: 0.25° Coverage: extended tropics, 1998-2010	NASA [110]
GADM v2	Administrative boundaries level 4: Village Development Committees (VDC) Coverage: global	www.gadm.org
FRIEND	Daily streamflow gauges Coverage: Nepal (50 gauges)	FRIEND [105, 66]
CENSUS	VDC population (Households) Coverage: Nepal	2011 Census of Nepal [54]

Table 6.1: Publicly available datasets.

economic performance metrics listed in table Table 6.3. The sensitivity of these features to uncertainties in local economic characteristics can be assessed using the slider panel. The panel also allows the user to toggle between two alternative streamflow regionalization approaches. The prediction performance each of these methods at existing stream gauges can be visualized on the map. Alternatively, the *Regional Mapper* mode allows the output metrics to be mapped nationally, with the objective of providing policy makers with an overview of the spatial distribution of mico-hydropower feasibility in Nepal. Users select a metric, and a chloropleth map of the median value for each community is displayed on the interactive map (Figure 6.2 (a)).

Symbol	Description	Default	Source
$\overline{kW_{HH}}$	Average electrical capacity per household	$0.1[kW]$	Chapter 5
$\overline{P_{kW}}$	Average monthly household expenses for (off-grid) electricity	$2[NRp]$	Chapter 5
γ_P	Price elasticity of (off-grid) electricity demand	-0.14	Chapter 5
α_0	Average cost of a 1-kW micro-hydro scheme	$207[kNRp]$	Chapter 5
α_P	Scale elasticity of micro-hydro costs	0.93	Chapter 5
p_{FI}	Feed in Tariff per kWh (grid-connected case)	$3.5[NRp]$	[18]*
η_0	Maximum turbine yield	0.51	[146]
α	Turbine cutoff flow	0.1	[146]
i	Discount rate	15[%]	[87, 122]
n	Infrastructure lifespan or loan payback period	$7[y]$	[87, 122]
FDC	Method used to estimate the flow duration curve at the intake	Müller Thompson Chalise	[153] [56]

Note: * 3.5NRp/kW is the current bulk purchase price for grid power in Nepal

Table 6.2: Local user inputs. Default values are national averages (see provided refereneces) and can be tuned locally by the user. If the user chooses to submit them, the updated local parameters parameters are reviewed and incorporated to the tool to generate more accurate regional feasibility maps.

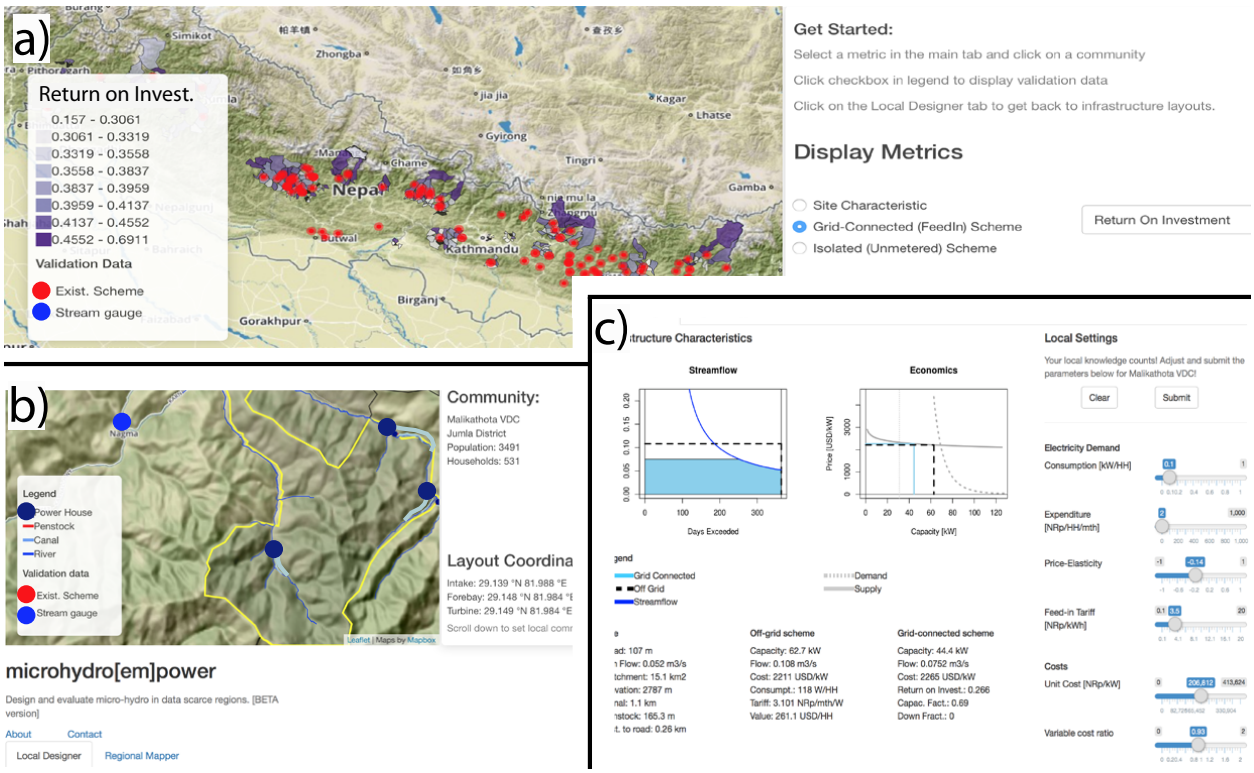


Figure 6.2: Screenshots of the webtool GUI: (a) Interactive map in *Regional Mapper* mode showing the return on investment expected from micro-hydro development in each VDC. Existing schemes are mapped as red dots. (b) Interactive map in *Local Designer* mode displaying infrastructure layouts in a selected community and a nearby stream gauge; (c) Performance metrics of the selected scheme and slider panel to tune and submit local economic parameters;

6.2.3 Computational Modules

The computational workhorse of the tool is hosted on a server and consists of four modules that interact (as illustrated in Figure 6.3) to optimize the placement and capacity of micro hydropower schemes, and evaluate their economic performance. The topography module uses a digital elevation model (DEM) to optimize the placement of micro hydro infrastructure near the considered community, as described in Part (a). Once the location of the scheme is determined, the hydrology and economics modules (Part (b)) use the methods developed in Chapters 2 to 5 to predict local constraints on the capacity of the micro hydro plant, which is optimized in the fourth module (Part (c)).

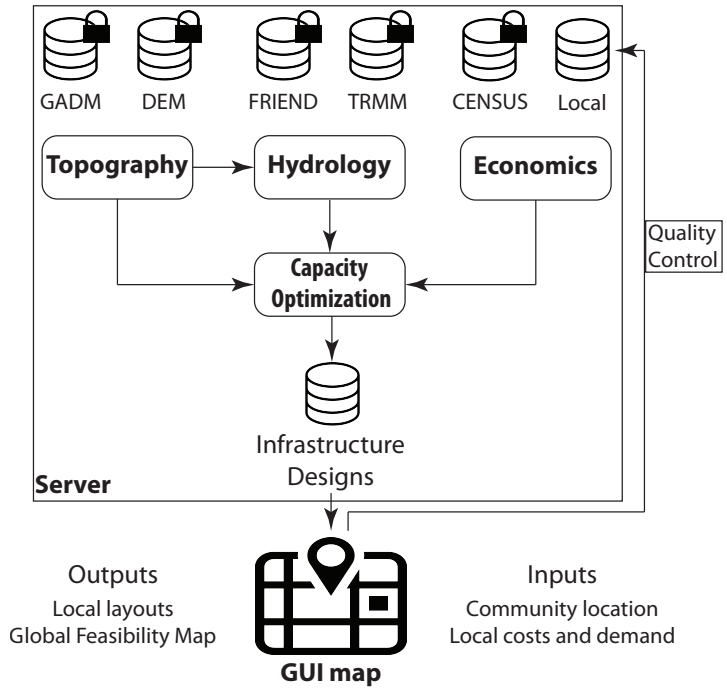


Figure 6.3: Webtool architecture. The four computational modules on the server (Section 6.2.3) are represented as rounded rectangles. Locked databases represent available data stored in the server (Table 6.1). Unlocked databases represent the local economic inputs and infrastructure design outputs updated on the server when the user interacts with the graphic user interface map. User-provided local parameters are reviewed and evaluated before being included in the tool.

(a) Topography

The location of micro hydro infrastructure components on the landscape determines key salient features, like costs, hydraulic head and mean flow, that drive the economic performance of the scheme, and must be accounted for when mapping hydropower potential.

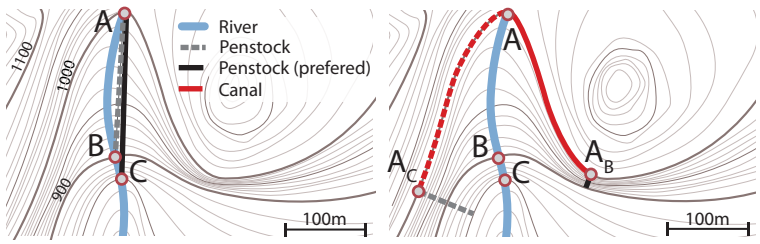
Infrastructure siting is a complex and typically heuristic design process based on numerous site visits and constrained by factors including technical, costs, accessibility, flood safety, legal and environmental considerations [e.g., 124, 60].

The process is driven by topography, which affects both the potentially harvested power, through the hydraulic head and the area of the contributing catchment; and the cost of the infrastructure, by affecting the lengths of the penstock and headrace canal. This dependence on topography allows layout optimization to be automated – albeit in a simplified version – for potential mapping purposes, thanks to the global availability of free, high resolution digital elevation models (DEM) from remote sensing platforms [e.g., 248, 135].

In their most basic form, existing potential mapping algorithms [17, 137] compute watershed boundaries and river reaches from a digital elevation model, and multiply the elevation difference obtained within each reach (or arbitrary river segment [130]) with the area of the corresponding catchment and a regional runoff parameter to estimate gross hydropower potential. In a more sophisticated approach, Yi et al. [248] implemented a cell-by-cell search algorithm along the drainage network. Unlike previous approaches, the method allows for water diversions from lower to higher order streams and identifies potential (straight line) waterways from each stream pixel within a series of predefined search radii. The hydropower potential of each waterway is then evaluated based on its average (straight distance) slope. A similar search algorithm was further developed by Larentis et al. [135], allowing for water storage reservoirs, and accounting for the effect of preexisting schemes on the exploitable potential. A common aspect of all these methods is their sole reliance on the elevation profile of the stream channel to evaluate the hydraulic head. Yet hillslope topography may also have a significant impact on infrastructure placement, as illustrated in Figure 6.4. Hillslopes are typically steeper than stream beds in mountainous regions, and the optimal penstock position may be located on a favorable slope at a significant distance from the stream. Headrace canals of up to two kilometers are commonly found in Nepal.

The Micro Hydro [em]Power tool explicitly includes hillslope topography in the optimization. The algorithm optimizes the position of the intake, forebay and power house of micro-hydropower schemes to maximize topographic suitability, defined as the product of the hydraulic head and the area of the contributing catchment. Generated layouts are restricted to technically feasible options using heuristic constraints. In line with anecdotal evidence from Nepal [180], the lengths of the headrace canal and penstock are capped at $2km$ and $200m$ respectively, and intakes with contributing catchment areas of less than $10km^2$ are rejected. The average slope of the penstock is also restricted to values between 0.176 and 1. The higher bound is suggested by Junejo et al. [125, p. 52] for reasons of constructability and slope stability. The lower bound is obtained by considering maximum friction losses of 10% of the net hydraulic head, as recommended in Chitrakar [60], and assuming a linear head loss coefficient of 0.016 (see Appendix E.1). The ensuing algorithm consists of the following steps, illustrated in Figure 6.5:

1. The first step generates a stream network using topographic information from the DEM. Slope, aspect and flow accumulation rasters are computed using the A^T search



(a) Layouts solely based on streambed topography **(b)** Layouts accounting for hillslope topography

Figure 6.4: Stylized example showing the importance of accounting for hillslope topography in the optimization of run-of-river hydropower layout. In subfigure (a), layouts are evaluated solely based on streambed topography: layout AC is preferable to layout AB because its has a steeper average slope. Subfigure (b) accounts for local hillslope topography by allowing water to be diverted (canals in red). Under these conditions, layout AB can now take advantage of the favorable slope conditions at point A_B and becomes preferable to layout AC

algorithm implemented in the r.watershed function in GRASS [79].

2. The second step identifies and excludes DEM raster cells that cannot contain the penstock. Valid cells are (i) within the administrative boundaries of the community, (ii) within 2km of a river, (iii) within the altitude range covered by the rivers in the community and (iv) have a slope between 0.176 and 1.

3. The third step optimizes the layout of the penstock. Its position and direction are determined by considering each valid DEM cell in decreasing order of slope and extracting an elevation profile along the local flow direction given by the aspect raster. Penstock length L along the elevation profile is determined by maximizing the net hydraulic head
$$H(L) = \Delta z(L) - k \cdot L$$
where $\Delta z(L)$ is the elevation difference along the penstock and $k = 0.016$ are the assumed linear friction losses. Penstock length L is capped at $200m$. The forebay and power house are located at the higher and lower end of the penstock respectively.

4. The headrace canal is determined as the DEM contour line running through the forebay, and the intake is positioned at the intersection of the contour line with the nearest stream. The layout is discarded if the length of the canal is larger than $2km$, or if the catchment area A upstream of the intake is larger than $10km^2$.

Steps 3 and 4 are iterated until the desired number of infrastructure layouts are found and returned in decreasing order of topographic suitability. By default, the tool generates five valid layouts per community and uses the characteristics of the median layout, in terms of topographic suitability, to construct the large scale feasibility map.

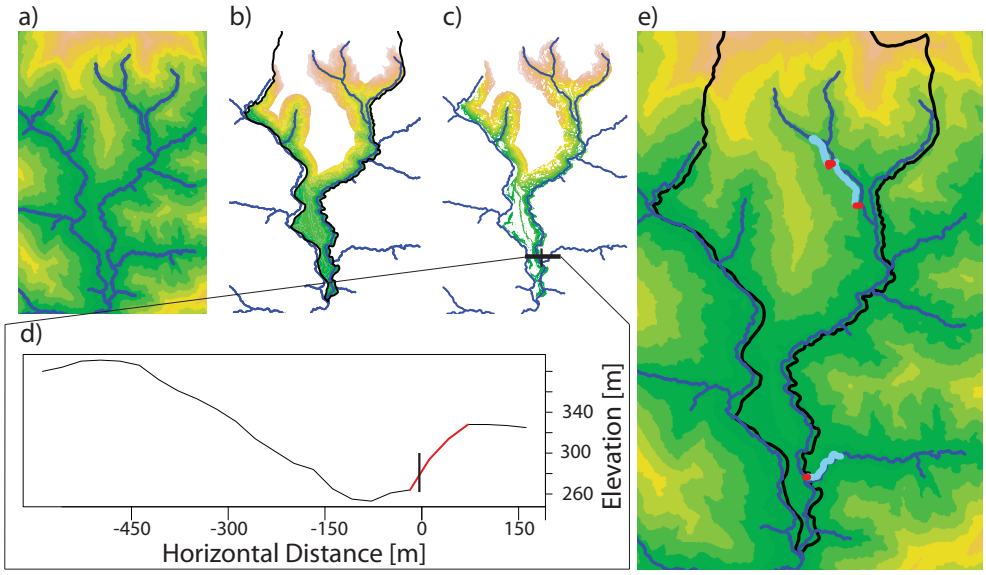


Figure 6.5: Topographic layout optimization algorithm: DEM analysis to determine the location of streams and generate slope and aspect rasters (a); Pixel selection based on elevation (b) and slope (c) criteria; Determination of optimal penstock location (c); Optimization of penstock length based on the elevation profile (d); Determination of the head race canal using DEM contours (e).

(b) Hydrology and economics

The hydrology module uses remotely sensed rainfall and topography to predict flow duration curves at the intake locations provided by the topography module. By default, the module uses the probabilistic process-based model described in Chapter 3, with the option of selecting the empirical model provided in Chalise et al. [56] instead. An ad hoc comparative analysis has shown that both methods have comparable (mean) cross-validation performances. The empirical method has shown less variability in its performance, but the process-based approach is expected to be more robust to a changing and variable climate. For both models, biases in remote sensing precipitation were adjusted as described in Chapter 2, and runoff variables were regionalized from stream gauges using the geostatistical approach described in Chapter 4.

The economics module uses cost, demand and financial parameters, either user inputs or default values from Chapter 5, to generate cost and demand curves for micro-hydropower electricity. These curves respectively represent the effect of infrastructure capacity kW (i.e. the size of the turbine) on its cost P, and the effect of the unit price of electrical capacity P_{kW} on the electrical capacity kW_{HH} consumed by households. Consistent with Chapter 5,

double-log functions are assumed for both curves:

$$\ln P = \ln \alpha_0 + \alpha_{kW} \ln kW \tag{6.1}$$

$$\ln kW_{HH} = \ln \gamma_0 + \gamma_P \ln P_{kW} \tag{6.2}$$

where α_0 represents the local costs of a 1-kW infrastructure and α_{kW} the related economies of scale, that is the relative decrease in average (unit) cost when doubling infrastructure size; γ_P is the price elasticity of demand that can be estimated at a local level, albeit with potentially significant biases [103], using contingent valuation surveys, where respondents are asked to state their willingness to pay for hypothetical consumptions of electricity; lastly, unit price consumption γ_0 can be directly computed using Equation (6.2) if local price elasticity, average consumption of electrical capacity and the associated unit price are known.

(c) Capacity optimization

The capacity optimization module uses outcomes from the three previous modules to perform an economic optimization of the infrastructure capacity and estimate the associated costs and economic performance metrics described in Table 6.3. For each optimization, the algorithm successively considers the case of grid-connected infrastructure with exogenous energy-based pricing (i.e. feed-in tariff) and off-grid electrification with endogenous capacity-based pricing (i.e. unmetered local households).

Grid-connected case The determination of the capacity flow Q_d of run of river hydropower (i.e. the maximum flow a plant can process) is a well characterized economic optimization problem driven by the cost function of the scheme and constrained by the variability of the available flow. With notable exceptions [22], the high dimensionality of the problem and the empirical nature of the typical flow duration curves has traditionally prescribed the use of data-intensive numerical optimization algorithms, which has been the focus of substantial research [see 101, for a complete review]. Most algorithms assume profit maximization and an exogenously determined retail price for each kWh of generated electricity. These assumptions are appropriate for privately-owned, grid-connected infrastructure benefitting from power purchase agreements with feed-in tariffs, but are not applicable to off-grid electrification, as discussed in Section 6.1. The following method is nonetheless included in the tool, with the purpose of anticipating the possibility of eventually feeding the generated electricity into the grid, in the event that the grid should reache the community.

Assuming ergodic daily streamflows, the annual energy production E_y can be expressed as a function of the capacity flow:

$$E_y = gH\rho \int_Y Q(D)\eta\left(\frac{Q(D)}{Q_d}\right) dD, \tag{6.3}$$

with g the gravitational acceleration, ρ the density of water and $Q(D)$ the flow duration curve. Following Basso and Botter [22], a single turbine with an efficiency curve that can be

represented as a step function is assumed:

$$\eta(Q) = \eta_0 \cdot \delta(Q \geq \alpha Q_d),$$

where $\delta(\cdot)$ is the Dirac delta, η_0 the maximum power yield of the turbine and α the cutoff flow. The tool determines the optimal capacity flow by maximizing the net present value of the infrastructure through numerical optimization:

$$\max_{Q_d}\{\hat{r} \cdot p \cdot E_y(Q_d) - P(Q_d)\}$$

where $\hat{r} = \frac{1}{r}\left(1 - \frac{1}{(1+r)^n}\right)$ is a multiplier that accounts for the project lifespan n and the considered discount rate r, p is the exogenously determined market price of electricity and P is the cost of the infrastructure provided by Equation (6.1).

Off-grid case Although prevalent in developing countries and very different from grid-connected systems, off-grid micro hydro infrastructure with unmetered household connections has received relatively little attention in the literature. The economic performance of these schemes is driven by the consumption of local users. Following their demand curves, households pay an agreed-upon fee for a chosen electrical capacity. In the absence of meters, households do not purchase energy units, but options on peak energy consumption. Grid connected schemes are part of a competitive market, where numerous power producers sell their output to the central power utility; price is therefore exogenous and as price takers, each power producer optimizes their capacity to maximize profit by equating marginal costs to marginal revenues. In contrast, a power utility managing off-grid power infrastructure is in a situation of natural monopoly and can determine their own price. In that situation, because of the inelastic nature of electricity demand (Chapter 5), price regulation mechanisms must be enforced to ensure equitable access to electricity. A price-inelastic demand implies negative marginal revenues for all levels of consumption. Because no level of production would allow marginal revenues to reach (positive) marginal costs, a profit-maximizing monopolist will boundlessly increase price and decrease output. In other words, unregulated electricity prices will cause the utility owner to produce minimal electrical output that they will sell to households paying the highest price.

As a result, many off-grid micro-hydropower schemes in Nepal are subject to some level of participative pricing [128, p. 36] and do not operate solely on a profit maximizing basis. Although numerous types of public pricing policies exist [16], the optimization used in Micro Hydro [em]Power considers average cost pricing, as recommended in Junejo [124, p. 61]. This implies that the price of the produced electricity is regulated so as to allow the full recovery of infrastructure costs, but no profit. This is consistent with the assumptions made in the econometric model described in Chapter 5.

Accordingly, the optimal capacity of off-grid microhydropower schemes is determined by the tool as follows. Cost recovery occurs when average costs are exactly compensated by

average revenues. In demand-limited cases, that is when streamflow is sufficient to generate enough electricity to satisfy demand, the optimal power capacity kW is obtained by solving

$$\alpha_0 kW^{\alpha_{kW}-1} = \left[\frac{kW}{\gamma_0 HH}\right]^{1/\gamma_P}, \tag{6.4}$$

where the left hand side represents the average costs obtained from Equation (6.1), and the right hand side the average revenues derived from Equation (6.2), with HH the number of supplied households. The optimal flow capacity is then computed as $Q_d = \frac{kW}{\eta_0 \rho g H}$. In flow limited cases, streamflow variability affects the optimization, in so far as no revenue is generated when the available flow does not allow the demanded capacity to be generated. Optimal capacity is obtained by (numerically) solving Equation (6.4), this time after multiplying the average revenue (left hand side) by $F(Q_d)$, the cumulative density function of streamflow (that is the complement of the flow duration curve) taken at Q_d. $F(Q_d)$ can be interpreted as the capacity factor of the plant, that is the fraction of time when the available flow allows it to function at full capacity.

Symbol	Description	Units
H	Net hydraulic head	$[m]$
L_p	Length of penstock	$[m]$
L_c	Length of headrace canal	$[m]$
TSI	Topographic suitability index	$[kW \cdot (mm/y)^{-1}]$
Q_d	Design flow capacity	$[m^3 s^{-1}]$
kW	Power capacity	$[kW]$
AC	Unit cost	$[USD \cdot kW^{-1}]$
CF	Capacity factor (Grid-connected case)	$[-]$
ROI	Return on Investment (Grid-connected case)	$[-]$
W_{HH}	Household consumption (Off-grid case)	$[W \cdot HH]$
p_{UM}	Unmetered tariff (Off-grid case)	$[NRp \cdot kW^{-1} \cdot Mth^{-1}]$
$CVI = W_{HH} \cdot p_{UM}$	Community value index	$[NRp \cdot HH^{-1} \cdot Mth^{-1}]$

Table 6.3: Characteristics estimated by the tool for each generated micro-hydro scheme.

6.3 Evaluation

6.3.1 Data and methods

The predictive performance of the tool is assessed using the Renewable Energy Data Book (REDB) [7] published by the Alternative Energy Promotion Center (AEPC) of the Government of Nepal. The dataset is a publicly available infrastructure census that provides the capacity and cost of 148 micro hydropower schemes that were subsidized by the AEPC

and commissioned between 2007 and 2011. Most infrastructure is community-owned and all schemes supply villages that do not have access to the centralized grid. The dataset provides the location of the schemes at the ward level[3], as shown in Figure 6.6 (a), but the exact position of the infrastructure elements is not available.

The resulting uncertainty on the layout of existing schemes precludes a rigorous validation of the ability of Micro Hydro [em]Power to site and size infrastructure with respect to local topography and demand. In fact, both design processes result from ill-defined and eminently local optimizations that are challenging to emulate in a large scale remote assessment tool. Here I use regression techniques and point pattern analysis to test the following three hypotheses:

H1 The tool identifies the most appropriate communities (VDCs) for micro hydropower development, as evidenced by the number of existing schemes in each community.

H2 The tool can successfully identify the most promising wards for micro hydropower within these communities.

H3 The tool predicts the position and capacity of existing micro hydropower schemes.

Statistical inference based on generalized linear models can be used to test hypothesis 1. The 108 communities possessing at least one existing scheme and their 488 immediate neighbors were sampled. This sampling strategy minimizes the potential effect of unobserved variables on the presence of a micro-hydro scheme. In particular, neighboring communities are assumed to have access to comparable information, technical expertise and loan conditions as those where micro hydro facilities are located. A Poisson regression was used to evaluate the tool's ability to predict the number of micro hydro schemes in the considered communities. By assuming a Poisson-distributed response variable, this form of regression analysis is adapted to model count data, here the number of micro hydro scheme per community. The independent (right hand side) variables were the three feasibility metrics provided by the tool based on different design assumptions. The Topographic Suitability Index (TSI) only considers the effect of topography on the gross extractable energy. The Return on Investment (ROI) ratio assumes a privately managed, profit maximizing micro hydro utility. The scheme is assumed connected to the centralized grid and benefitting from feed in tariffs on the injected electricity. Unlike the TSI, the ROI is constrained by infrastructure costs. Finally, the Community Value Index (CVI) assumes an off-grid micro-hydro scheme, managed by a local public utility with the objective of recovering costs (without generating profits). The CVI is constrained by local community demand and is perhaps the most realistic metric in the current situation. Student's t-tests on the regression coefficients were used to evaluate the significance of the relation between these three metrics and the presence and number of micro hydro schemes in the sampled communities.

[3]The ward is the smallest administrative subdivision in Nepal. There are 9 wards in a village development committee (VDC).

Hypothesis 2 was tested by assessing whether the tool improved on the prediction of the location of existing micro hydro schemes within the communities, compared to an alternate method, where predictions were generated randomly. The set of predicted and observed micro hydro locations was modeled as a marked random point process. An extension of Ripley's K [187] was used to evaluate the statistical significance of clusters between predicted and observed locations. The K-function was modified to allow for multiple supports because the tool optimizes the location of micro hydro sites independently for each community: predicted sites can only be compared to actual sites within a community. Individual K-functions must therefore be estimated independently for each community and aggregated across supports. The construction of the aggregated estimator and its use in a Monte Carlo test for statistical inference are detailed in Appendix E.2.

Linear regressions were finally used to address the third hypothesis. In contrast to hypothesis 1, the 148 existing schemes (and not the communities containing them) were sampled with the objective of testing the tool's ability to predict the optimal capacity and layout of micro hydropower plants. Student's t-tests on the resulting ordinary least square coefficients were used to evaluate the significance of the relation between the median infrastructure capacities predicted by the tool (both under the grid-connected and off-grid assumptions) for each community, and the actual capacity of the corresponding existing schemes. Two important caveats must be considered when interpreting regression results. First, the available data do not allow a formal test of the tool's predictions about the optimal position of infrastructure. The regression results should nonetheless be sensitive to the position of the infrastructure because optimal capacity is driven by inherently local topographic effects, as elaborated in Section 6.2.3(a). A good prediction of micro-hydro capacity is likely associated with a good prediction about its position. Second, the intended purpose of the tool is not to emulate *existing* schemes, as tested by hypothesis 3, but rather to optimize the layout and design of micro-hydro components near a given community. Thus, the fit of the linear regression is not indicative of the tool's ability to fulfill purpose, but may rather indicate the presence of unaccounted factors influencing the design of existing schemes.

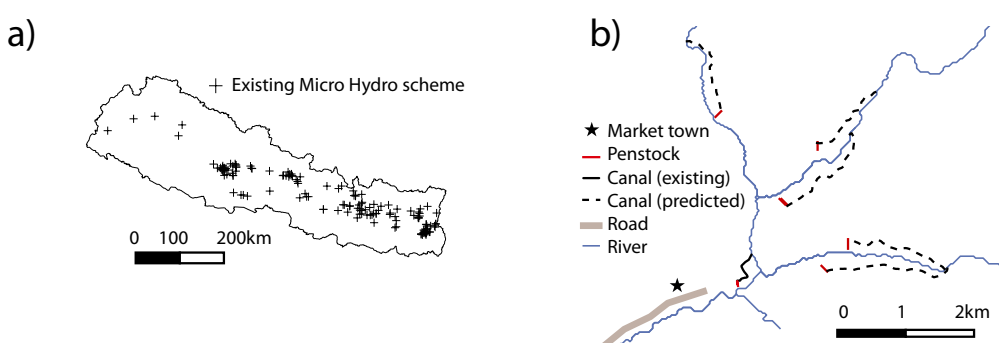


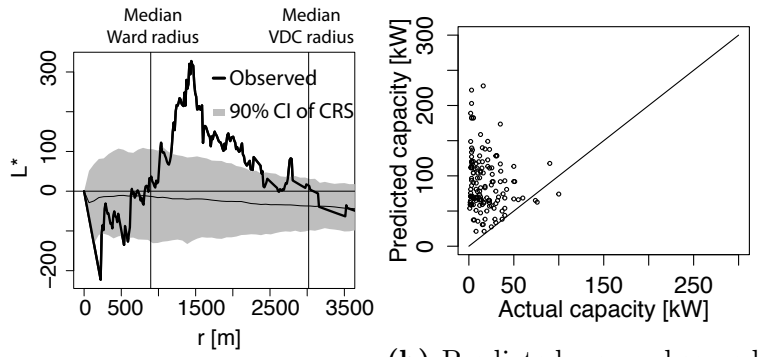
Figure 6.6: Validation data. (a) Approximate location of the 148 microhydro schemes of the Renewable Energy Data Book [7]; (b) Layout of the existing micro-hydro scheme on the Lohore Kohla (solid) and infrastructure positions predicted by the tool (dashed).

6.3.2 Results and discussion

Evidence suggests that the Micro Hydro em[Power] effectively identifies promising locations for micro hydro development. The significantly positive association found between the community value index (CVI) and the number of existing schemes indicates that the tool successfully identifies communities, where micro hydro is particularly promising. Results of the Poisson regression (Table 6.4 (1)) show that the number of schemes per VDC is expected to increase by 1.2% for each marginal unit of CVI estimated by the tool. The estimated coefficients also indicate that neither the topographic suitability index (TSI), nor the return on investment ratio (ROI) are significantly correlated to the number of schemes. Unlike the CVI, neither metric accounts for electricity demand, which illustrates the role of local demand, and particularly the size of the local community, as a limiting factor for micro hydro feasibility. The point pattern analysis further suggests that the tool reliably predicts suitable zones for micro-hydro development *within* communities. The normalized empirical K-function shown in Figure 6.7 (a) indicates that predicted locations are clustered around actual micro-hydro schemes in a statistically significant manner over distances smaller than the median radius of the considered communities (3km). The clustering effect disappears at distances below 1km, which roughly corresponds to the median radius of wards and is consistent with the expected uncertainties on the location of existing schemes, which are approximated at the centroid of the wards.

	Dependent variable:	
	$N_{MHP_{REDB}}$	kW_{REDB}
	Poisson	*OLS*
	(1)	(2)
TSI	−0.030 (0.029)	
ROI	−2.061 (1.627)	
CVI	0.012** (0.005)	
kW_{Grd}		−0.0003 (0.002)
kW_{OffGrd}		−0.047* (0.026)
Constant	−3.970*** (1.235)	21.843*** (2.835)
Observations	597	149
Adjusted R^2		0.009
Note:	*p<0.1; **p<0.05; ***p<0.01	

Table 6.4: Validation of the webtool on the REDB dataset. Specification (1) represents the tool's ability to predict the number of micro hydro schemes on the REDB communities and their immediate neighbors. The statistically significant coefficient for community value index (CVI) suggests that local electricity demand is a strong indicator of micro hydropower feasibility. Specification (2) shows the tool's performance in predicting the capacity of the existing micro hydro schemes listed in the REDB.



(a) Modified Ripley's K

(b) Predicted vs. observed capacity per scheme.

Figure 6.7: (a) Ripley's cross-K function is adapted to accommodate multiple support areas, as described in Appendix E.2, and normalized as $L^*(r) = \sqrt{K(r)/\pi} - r$ to represent deviations from complete spatial randomness represented by the horizontal line $L^*(r) = 0$. The 90% confidence interval around the CRS was generated through Monte Carlo (N=1000). The position of the empirical curve above the confidence interval indicates significant clustering between actual and predicted micro-hydro sites for distances approximatively ranging between the median ward radius (0.9km) and the median VDC radius (3km). (b) The capacity of the observed micro hydro schemes is plotted against the median capacity predicted by the tool in the corresponding community, assuming off-grid electrification. The plot clearly shows that the tool has a strong tendency to overestimate the capacity of the existing schemes.

Micro Hydro [em]Power does not successfully predict the capacity of existing schemes, as evidenced by the poor fit ($R^2 < 0.01$) and negative or insignificant coefficients of the linear regression (Table 6.4 (2)). While indicative of a poor ability to emulate existing schemes, these results are not indicative of the tool's ability to fulfill its purpose: assisting communities in the identification of sites for new installations. The poor fit of the linear regression may instead suggest that the capacity of existing schemes is constrained by local factors that are not accounted for in the tool, as evidenced by its strong tendency to overestimate micro hydro capacity (Figure 6.7 (b)).

As an example of such local effects, consider the case of the Lohore Kohla power plant, a 23 kW micro hydropower scheme supplying 312 households in and around the rural market town of Namaule in Western Nepal. The existing scheme, which GPS coordinates were recorded in 2012, does not correspond to any of the layouts predicted by the tool, as seen on Figure 6.6 (b). Although topographically more advantageous, all predicted sites are located on the left bank of the Lohore Khola: their construction would require a new bridge to transport materials from the road, which would significantly increase their cost. In addition to road accessibility, plant proximity to a demand center can play a significant role in the placement of micro hydro schemes. This is true for Lohore Kohla, a fraction of the retrieved energy is directly used as mechanical power during the day for grain milling. While the

infrastructure itself is owned by the community, the attached mill belongs to the operator of the scheme. This arrangement provides incentives for the operator to properly maintain the system, but requires the mill to be easily accessible from the market town to generate profits.

Although particularly relevant in Nepal, site accessibility and institutional arrangements are two key local factors constraining the design of micro hydropower infrastructure. These factors are challenging to assess *a priori*. They are often ill-defined and represent an inherent source of uncertainty in large scale assessment tools. The interactive nature of Micro Hydro [em] Power allows local demand, which appears to be a key constraining factor (see Table 6.4 (1)), to be assessed. The tool does not currently allow to account for the accessibility contraints of the Lohore Kohla case, but recent development in participatory GIS platform [e.g., 102] offer considerable scope for future improvement. Participatory GIS uses interactive web maps to create, assemble and disseminate geographic information provided voluntarily by individuals [93]. These capabilities can be used in future implementations of the tool better incorporate local qualitative constraints, for example by allowing users to constrain the topographic optimization algorithm to specific zones within their community.

6.4 Conclusion

This chapter presents Micro Hydro [em]Power, a web-based, open-source application to design and assess the feasibility of micro-hydropower for rural electrification. The tool leverages the increasing availability of information and communication technologies to empower local communities with the means to plan and assess micro-hydropower options in a quantitative way. Its interactive nature also allows cost and demand information to be assessed at the local level, in order to map the potential for micro hydropower more accurately at a regional scale.

The predictive performance of Micro Hydro [em]Power was evaluated against 148 existing schemes, showing its ability to identify promising communities and spot regions within these communities that are most favorable for micro hydropower development. The analysis also illustrates the importance of local constraints that must be assessed at the local level, as illustrated by the tool's ability to incorporate user-provided cost and demand information. Local constraints are often ill-defined and challenging to incorporate in a transferable design tool. Ultimately, Micro Hydro [em]Power will not replace proper participative planning and field-informed engineering design. Its purpose is rather to initiate the process by bridging the information gap between local knowledge to technical expertise in data-scarce regions.

Chapter 7

Conclusion

A brief summary of the key findings is presented to close this dissertation. These conclusions underline a number of avenues for future development that are also described below.

7.1 Summary of Findings

The purpose of this dissertation was to develop tools and methods to assist the design and evaluation of micro hydropower infrastructure in data-scarce regions. The research leverages globally available data to address a major information gap between technical experts who *design* infrastructure, and community members who *use* it. Such information barriers explain the poor penetration of micro hydropower in Nepal despite very favorable conditions, as argued in Chapter 1.

Chapter 2 addressed the challenge of estimating precipitation over large spatial areas. This is an ongoing challenge for hydrologists because sparse rain gauges networks do not provide a robust basis for interpolation, and the reliability of remote sensing products, although improving, is still imperfect. I proposed a novel procedure to combine these two sources of observation to accurately predict spatially distributed rainfall. Unlike existing bias adjustment techniques, the algorithm corrected the possibly confounding effects of different sources of errors in satellite estimates by adjusting biases on rainfall *frequency* in addition to rainfall intensity. It explicitly accounted for the spatial heterogeneity of the biases, and made full use of sparse and non overlapping historical ground observations. The method was applied to adjust the distributions of daily rainfall observed by the TRMM satellite in Nepal, which exemplifies the challenges associated with a sparse gauge network and large biases due to complex topography. I found that using a small subset of the available gauges, the proposed method outperformed alternate approaches that use the complete network of available gauges to either directly interpolate local rainfall, or correct TRMM by adjusting monthly means.

Chapter 3 addressed the challenge of evaluating flow duration curves in sparsely gauged regions, where flow observations are missing to construct them empirically. Flow duration

curves are critical inputs to the design of micro hydropower and provide important information on the variation in the availability and reliability of surface water to supply ecosystem services and satisfy anthropogenic needs. I used a probabilistic approach to derive an analytical expression for flow duration curves in seasonally dry climates. During the wet season, streamflow was modeled as a stochastic process driven by rainfall and governed by soil moisture dynamics and subsurface storage. During the dry season, streamflow was modeled as a deterministic recession with a stochastic initial condition that accounts for the carryover of catchment storage across seasons. The resulting analytical expression was applied in Nepal, coastal California and Western Australia, where it successfully modeled flow duration curves using five physically meaningful parameters with minimal calibration. The approach successfully produced period-of-record flow duration curves and allowed inter-annual and intra-annual sources of variations to be separated.

Chapter 4 addressed the challenge of interpolating runoff variables on stream networks. While most parameters of the flow duration curve model are readily obtainable from remote sensing data, a few variables relate to runoff processes taking place below the land surface and not detectable by space-borne platforms. To address this issue, I introduced TopREML as a method to predict runoff variables in ungauged basins using streamflow observations from nearby gauges. The approach used a linear mixed model with spatially correlated random effects and a restricted maximum likelihood estimator. The nested nature of streamflow networks was taken into account by using water balance considerations to constrain the covariance structure of runoff and to account for the stronger spatial correlation between flow-connected basins. The method was successfully tested in cross validation analyses on mean streamflow and runoff frequency in Nepal (sparsely gauged) and Austria (densely gauged), where it matched the performance of comparable methods in the prediction of the considered runoff signature, while significantly outperforming them in the prediction of the associated modeling uncertainty. TopREML's ability to combine deterministic and stochastic information makes it a particularly versatile method that can readily be applied in both densely gauged basins, where it takes advantage of spatial covariance information, and data-scarce regions, where it can rely on covariates, which are increasingly observable thanks to remote sensing technology.

Chapter 5 described an econometric approach to estimate the cost function and demand curve of off-grid micro hydropower in Nepal. The empirical strategy used a digital elevation model and population census to construct instrumental variables to estimate the respective effects of electricity price on consumption, and of infrastructure size on its cost. The method extracted cost and consumption data from an inventory of existing micro hydro infrastructure, making it well suited to the context of rural electrification in developing countries, where metered household consumption data are typically unavailable. Despite large standard errors, the resulting estimates were in line with anecdotal evidence from case studies in rural Nepal, but diverged from values typically found in previous studies in grid-connected regions. The analysis attributed these discrepancies to the notable influence on the estimated elasticities of two fundamental features of off-grid electrification in developing countries: in-kind community contributions to infrastructure construction and very low levels of household

electricity consumption.

Lastly, Chapter 6 presented Micro Hydro [em]Power, a computer tool assembling the methods developed in Chapters 2 to 5 along with a search algorithm using a digital elevation model to optimize the location of micro hydropower schemes. Unlike existing approaches, the algorithm accounts for the effect of local hillslope topography in the placement of infrastructure components. The tool is an open source application that can be accessed and operated on a web-browser (`http://mfmul.shinyapps.io/mhpower`). Its purpose is to assist local communities in the design and evaluation of micro hydropower alternatives in their locality, while using cost and demand information provided by local users to generate accurate feasibility maps at the national level, thus bridging the information gap. The predictive performance of Micro Hydro [em]Power was evaluated against 148 existing schemes in Nepal, showing its ability to identify promising communities and spot regions within these communities that are most favorable for micro hydropower development. The analysis also illustrated the importance of local design constraints. The tool accounts for some of these constraints by integrating user-provided local cost and demand information, but unaccounted local design constraints still represent a substantial source of uncertainty of the tool.

7.2 Future Work

Micro Hydro [em]Power is currently online and accessible to anybody equipped with a web browser. Once translated into Nepalese, the tool will be readily available to assist in the design and evaluation of micro hydropower in Nepal. There is nonetheless substantial scope for further development to improve the robustness and versatility of the tool.

First, the flow duration curve model derived in Chapter 3 restricts tool's use to rain-fed catchments in seasonally dry climates. The model remains to be extended to alpine and arid climates, where temperature is a significant stochastic driver of streamflow. Although incidentally not significant in Nepal, the effect of snow and glacial melt may be of particular relevance in mountainous regions where micro hydropower is most promising. The flow duration curve model has recently been extended to incorporate episodic snowmelt [195], but seasonal melt generates a substantial fraction of base flow in alpine regions and remains to be considered in probabilistic flow distribution models.

Second, substantial financial risks arise from stochastic variations in annual flow duration curves. An excessively long spell of dry years, when revenues from the produced electricity are lower than expected, will cause liquidity concerns that may jeopardize the financial sustainability of the micro hydro infrastructure. This risk is exacerbated by climate change, as shifts in the intensity and timing of daily streamflow have a direct impact on hydropower generation. The probabilistic character of the flow duration curve model allows (stationary) inter-annual streamflow variations to be isolated to predict financial risk (see case study in Chapter 3), which can be accounted for in the design of the infrastructure. Thanks to its process based nature, the model also offers interesting prospects to incorporate climate change by forcing a pseudo stationary stochastic rainfall input. These avenues are currently

being explored in the context of climate change attribution in Nepal.

Third, as alluded to in Chapter 6, substantial scope for improvement remains in the tool's ability to assess local design constraints. In particular, participatory GIS can be used to allow users to map features (e.g. topographic obstacles or legal considerations) constraining the placement of micro hydropower infrastructure in their locality. These restrictions can be used by the tool to limit the topographic search algorithm to specific zones within their community, where micro hydropower is feasible. Quantitative local constraints may be more challenging to obtain if some technical background is necessary to assess them. For instance, contingent valuation surveys used to estimate the price elasticity of demand are prone to considerable biases if not administered by a trained enumerator [140]. These difficulties warrant the need for a robust quality control mechanism to detect and correct such biases. Ultimately, each community is unique. Local constraints will always be an inherent source of uncertainty in large scale assessment tools, which will never be appropriate substitutes for proper participative planning and field-informed engineering design.

Bibliography

[1] 2015. URL http://www.worldbank.org/en/country/nepal/overview. 2, 3, 5

[2] MC Acreman and Michael J Dunbar. Defining environmental river flow requirements? a review. *Hydrology and Earth System Sciences*, 8(5):861–876, 2004. 35

[3] V Agafonkin. Leaflet: an open-source javascript library for mobile-friendly interactive maps, 2013. 106

[4] George A Aggidis, E Luchinskaya, R Rothschild, and DC Howard. The costs of small-scale hydro power production: Impact on the development of existing potential. *Renewable Energy*, 35(12):2632–2638, 2010. 84, 85, 89, 95

[5] Claudio A Agostini and Eduardo Saavedra. Elasticities of residential electricity demand in chile. Technical report, 2014. 97

[6] Chris M Alaouze. Reservoir releases to uses with different reliability requirements1, 1989. 34

[7] Alternative Energy Promotion Centre (AEPC). Renewable energy data book. Technical report, Governement of Nepal, 2009. 85, 116, 118

[8] Alternative Energy Promotion Centre (AEPC). Subsidy policy for renewable (rural) energy. Technical report, Governement of Nepal, 2009. 165

[9] Alternative Energy Promotion Centre (AEPC). Renewable energy data book. Technical report, Governement of Nepal, 2011. 85

[10] Takeshi Amemiya. The maximum likelihood and the nonlinear three-stage least squares estimator in the general nonlinear simultaneous equation model. *Econometrica: Journal of the Econometric Society*, pages 955–968, 1977. 92

[11] M. Andaroodi and A. Schleiss. Standardization of civil engineering works of small high-head hydropower plants and development of an optimization tool. *Communications LCH*, 26, 2006. 84, 85, 89

[12] C. Andermann, S. Bonnet, and R. Gloaguen. Evaluation of precipitation data sets along the Himalayan front. *Geochemistry Geophysics Geosystems*, 12(7):Q07023, 2011. doi: 10.1029/2011GC003513. 8, 22

[13] Christoff Andermann, Laurent Longuevergne, Stéphane Bonnet, Alain Crave, Philippe Davy, and Richard Gloaguen. Impact of transient groundwater storage on the discharge of himalayan rivers. *Nature geoscience*, 5(2):127–132, 2012. 34, 36, 39, 56

[14] A.M. Anders, G.H. Roe, B. Hallet, D.R. Montgomery, N.J. Finnegan, and J. Putkonen. Spatial patterns of precipitation and topography in the Himalaya. *Special Papers - Geological Society of America*, 398:39, 2006. doi: 10.1130/2006.2398(03). 22, 80

[15] J.D. Angrist and J.S. Pischke. *Mostly harmless econometrics: an empiricist's companion*. Princeton Univ Pr, 2008. 92, 95, 99

[16] W. Apgar and H.J. Brown. *Microeconomics and Public Policy*, chapter Chater 16 "Public Pricing". Scott Foresman and Company, 1987. 90, 115

[17] A Ballance, D Stephenson, R Chapman, and J Muller. A geographic information systems analysis of hydro power potential in south africa. *Journal of Hydroinformatics*, 2:247–254, 2000. 111

[18] Suraj Baral, Suman Budhathoki, and H Prasad Neopane. Grid connection of micro hydropower, mini grid initiatives and rural electrification policy in nepal. In *Sustainable Energy Technologies (ICSET), 2012 IEEE Third International Conference on*, pages 66–72. IEEE, 2012. 83, 108

[19] Pranab Bardhan and Christopher Udry. *Development microeconomics*. Oxford University Press, 1999. 88, 164

[20] Olga Barron, Mike Donn, SL Furby, Joanne Chia, and Chris Johnstone. *Groundwater contribution to nutrient export from the Ellen Brook catchment*. CSIRO Water for a Health Country, 2009. 53, 57

[21] A.P. Barros, M. Joshi, J. Putkonen, and D.W. Burbank. A study of the 1999 monsoon rainfall in a mountainous region in central Nepal using TRMM products and rain gauge observations. *Geophysical Research Letters*, 27(22):3683–3686, 2000. doi: 10.1029/2000GL011827. 22

[22] S. Basso and G. Botter. Streamflow variability and optimal capacity of run-of-river hydropower plants. *WATER RESOURCES RESEARCH,*, 48, 2012. 34, 50, 57, 114

[23] Nandita B Basu, P Suresh C Rao, Sally E Thompson, Natalia V Loukinova, Simon D Donner, Sheng Ye, and Murugesu Sivapalan. Spatiotemporal averaging of in-stream solute removal dynamics. *Water Resources Research*, 47(10), 2011. 57

[24] Michael Beenstock, Ephraim Goldin, and Dan Nabot. The demand for electricity in israel. *Energy Economics*, 21(2):168–183, 1999. 84, 88, 164

[25] Claude JP Bélisle. Convergence theorems for a class of simulated annealing algorithms on rd. *Journal of Applied Probability*, pages 885–895, 1992. 44, 159

[26] Luis Bendezú and Jose Gallardo. Econometric analysis of household electricity demand in peru, 2006. 98

[27] B.C. Bhatt and K. Nakamura. Characteristics of monsoon rainfall around the Himalayas revealed by TRMM precipitation radar. *Monthly Weather Review*, 133(1): 149–165, 2005. doi: 10.1175/MWR-2846.1. 21, 22

[28] Gary D Bishop and M Robbins Church. Automated approaches for regional runoff mapping in the northeastern united states. *Journal of Hydrology*, 138(3):361–383, 1992. 59

[29] R Bivand. spgrass6: Interface between grass 6 and r. *R package version 0.6-19*, 2010. 106

[30] Roger Bivand and Colin Rundel. rgeos: interface to geometry engine-open source (geos). *R package ver. 0.3–3.< http://cran. r-project. org/web/packages/rgeos/index. html*, 2013. 106

[31] Claudio JC Blanco, Yves Secretan, and André L Amarante Mesquita. Decision support system for micro-hydro power plants in the amazon region under a sustainable development perspective. *Energy for sustainable development*, 12(3):25–33, 2008. 105

[32] Günter Blöschl, Murugesu Sivapalan, Thorsten Wagener, Alberto Viglione, and Hubert Savenije. *Runoff Prediction in Ungauged Basins: Synthesis across Processes, Places and Scales*. Cambridge University Press, 2013. 2, 59, 105

[33] B. Bookhagen and D.W. Burbank. Topography, relief, and TRMM-derived rainfall variations along the Himalayas. *Geophysical Research Letters*, 33(8):L08405, 2006. doi: 10.1029/2006GL026037. 8

[34] DD Bosch, JM Sheridan, and FM Davis. Rainfall characteristics and spatial correlation for the georgia coastal plain. *Transactions of the ASAE*, 42(6):1637–1644, 1999. 80

[35] Ranjan Kumar Bose and Megha Shukla. Elasticities of electricity demand in india. *Energy Policy*, 27(3):137–146, 1999. 83, 97

[36] R.K. Bose and M. Shukla. Electricity tariffs in india: an assessment of consumers ability and willingness to pay in gujarat. *Energy Policy*, 29(6):465–478, 2001. 85

[37] G. Botter, F. Peratoner, A. Porporato, I. Rodriguez-Iturbe, and A. Rinaldo. Signatures of large-scale soil moisture dynamics on streamflow statistics across us climate regimes. *Water Resources Research*, 43(11):W11413, 2007. 35

[38] G. Botter, A. Porporato, I. Rodriguez-Iturbe, and A. Rinaldo. 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):2417, 2007. doi: 10.1029/2006WR005043. 9, 11, 35, 36, 37, 38, 39, 47, 81

[39] Gianluca Botter, Amilcare Porporato, Ignacio Rodriguez-Iturbe, and Andrea Rinaldo. Nonlinear storage-discharge relations and catchment streamflow regimes. *Water resources research*, 45(10), 2009. 35

[40] Gianluca Botter, Stefano Basso, Ignacio Rodriguez-Iturbe, and Andrea Rinaldo. Resilience of river flow regimes. *Proceedings of the National Academy of Sciences*, 110 (32):12925–12930, 2013. 35

[41] John Bound, David A Jaeger, and Regina M Baker. Problems with instrumental variables estimation when the correlation between the instruments and the endogenous explanatory variable is weak. *Journal of the American statistical association*, 90(430): 443–450, 1995. 99

[42] J. Briscoe. The financing of hydropower, irrigation and water supply infrastructure in developing countries. *International Journal of Water Resources Development*, 15(4): 459–491, 1999. 1

[43] F.P. Brissette, M. Khalili, and R. Leconte. Efficient stochastic generation of multi-site synthetic precipitation data. *Journal of Hydrology*, 345(3):121–133, 2007. doi: 10.1016/j.jhydrol.2007.06.035. 11

[44] Wilfried Brutsaert and John L Nieber. Regionalized drought flow hydrographs from a mature glaciated plateau. *Water Resources Research*, 13(3):637–643, 1977. 40, 46

[45] S Budhathoki. *Grid Connection of Gotikhel Micro Hydropower Plant Without Interrupting Isolated Load*. GRIN Verlag, 2011. 83

[46] Bureau of Meteorololgy. *Climate Data Online*. Government of Australia, 2013. 48

[47] California Irrigation Management Information System. *Daily Rainfall Data*. Department of Water Resources, 2013. 48

[48] Christian E Casillas and Daniel M Kammen. The delivery of low-cost, low-carbon rural energy services. *Energy Policy*, 39(8):4520–4528, 2011. 83

[49] A. Castellarin, G. Galeati, L. Brandimarte, A. Montanari, and A. Brath. Regional flow-duration curves: reliability for ungauged basins. *Advances in Water Resources*, 27(10):953–965, 2004. 46, 57

[50] A. Castellarin, R. Vogel, and A. Brath. A stochastic index flow model of flow duration curves. *Water Resources Research*, 40(3):W03104, 2004. 42

[51] A Castellarin, G. Botter, DA Hughes, S. Liu, T.B.M.J. Ouarda, J Parajka, D. Post, M Sivapalan, C. Spence, A. Viglione, and R. Vogel. *Prediction of flow duration curves in ungauged basins, Chp. 7 in Runoff Prediction in Ungauged Basins: Synthesis across Processes, Places and Scales (Eds. G. Blschl, M. Sivapalan, T. Wagener, A. Viglione, H. Savenije)*. Cambridge University Press, 2013. 34, 35, 57

[52] Attilio Castellarin, Giorgio Camorani, and Armando Brath. Predicting annual and long-term flow-duration curves in ungauged basins. *Advances in Water Resources*, 30 (4):937–953, 2007. 42

[53] S Castiglioni, A Castellarin, A Montanari, JO Skøien, G Laaha, and G Blöschl. Smooth regional estimation of low-flow indices: physiographical space based interpolation and top-kriging. *Hydrology and Earth System Sciences*, 15(3):715–727, 2011. 61

[54] Central Bureau of Statistics. National population and housing census. *Government of Nepal*, 2011. 105, 107

[55] Serena Ceola, Gianluca Botter, Enrico Bertuzzo, Amilcare Porporato, Ignacio Rodriguez-Iturbe, and Andrea Rinaldo. Comparative study of ecohydrological streamflow probability distributions. *Water Resources Research*, 46(9), 2010. 35, 46, 51

[56] S.R. Chalise, S.R. Kansakar, G. Rees, K. Croker, and M. Zaidman. Management of water resources and low flow estimation for the himalayan basins of nepal. *Journal of Hydrology*, 282(1-4):25–35, 2003. 57, 70, 108, 113

[57] M.J.M. Cheema and W.G.M. Bastiaanssen. Local calibration of remotely sensed rainfall from the TRMM satellite for different periods and spatial scales in the Indus Basin. *International Journal of Remote Sensing*, 33(8):2603–2627, 2012. doi: 10.1080/01431161.2011.617397. 8, 9

[58] Y Chen, E.E. Ebert, K.J.E. Walsh, and N.E. Davidson. Evaluation of TRMM 3b42 precipitation estimates of tropical cyclone rainfall using PACRAIN data. *Journal of Geophysical Research: Atmospheres*, 2013. doi: 10.1002/jgrd.50250. 8, 27

[59] Lei Cheng, Mary Yaeger, A Viglione, Evan Coopersmith, Sheng Ye, and Murugusu Sivapalan. Exploring the physical controls of regional patterns of flow duration curves–part 1: Insights from statistical analyses. *Hydrology and Earth System Sciences*, 16 (11):4435–4446, 2012. 35

[60] Pushpa Chitrakar. Micro-hydropower design aids manual. *SHPP/GTZ-MGSP/AEPC*, 2004. 4, 86, 111

[61] Ven Te Chow. Handbook of applied hidrology: a compendium of water-resources technology. In *Handbook of applied hidrology: a compendium of water-resources technology*. McGraw-Hill, 1964. 34

[62] Evan Coopersmith, MA Yaeger, Sheng Ye, Lei Cheng, and Murugusu Sivapalan. Exploring the physical controls of regional patterns of flow duration curves–part 3: A catchment classification system based on regime curve indicators. *Hydrology and Earth System Sciences*, 16(11):4467–4482, 2012. 35

[63] Robert R Corbeil and Shayle R Searle. Restricted maximum likelihood (reml) estimation of variance components in the mixed model. *Technometrics*, 18(1):31–38, 1976. 67

[64] N Cressie. *Statistics for Spatial Data*. Wiley-Interscience, New York, 1993. 60, 61, 62, 66, 77, 156

[65] Noel Cressie, Jesse Frey, Bronwyn Harch, and Mick Smith. Spatial prediction on a river network. *Journal of Agricultural, Biological, and Environmental Statistics*, 11(2): 127–150, 2006. 61

[66] Department of Hydrology and Meteorology. *Daily Streamflow and Precipitation Data*. Kathmandu, 2011. 22, 48, 69, 107

[67] P. Dhungel. *Financial and Economic Analysis of Micro-Hydro Power in Nepal*. case, The University of Minnesota, 2009. 85, 86, 165

[68] Romesh Dias-Bandaranaike and Mohan Munasinghe. The demand for electricity services and the quality of supply. *The Energy Journal*, 4(2):49–71, 1983. 83, 88, 164

[69] T. Dinkelman. The effects of rural electrification on employment: New evidence from south africa. *Quarterly Journal of Economics*, 2009. 84, 85

[70] Philip M Dixon. Ripley's k function. *Encyclopedia of environmetrics*, 2002. 168

[71] Paolo D'Odorico and Riccardo Rigon. Hillslope and channel contributions to the hydrologic response. *Water resources research*, 39(5), 2003. 65

[72] F Dominguez, E Rivera, DP Lettenmaier, and CL Castro. Changes in winter precipitation extremes for the western united states under a warmer climate as simulated by regional climate models. *Geophysical Research Letters*, 39(5), 2012. 34

[73] Wouter Drinkwaard, Arjan Kirkels, and Henny Romijn. A learning-based approach to understanding success in rural electrification: insights from micro hydro projects in bolivia. *Energy for Sustainable Development*, 14(3):232–237, 2010. 104

[74] Jeffrey A Dubin. *Consumer durable choice and the demand for electricity*, volume 1. Elsevier, 2014. 91

[75] Esther Duflo and Rohini Pande. Dams. *The Quarterly Journal of Economics*, 122(2): 601–646, 2007. 84, 85

[76] J. Duncan and E.M. Biggs. Assessing the accuracy and applied use of satellite-derived precipitation estimates over Nepal. *Applied Geography*, 34:626–638, 2012. doi: 10. 1016/j.apgeog.2012.04.001. 9, 28

[77] Earth Resources Observation and Science (EROS) Center. *HYDRO1k Elevation Derivative Database*. U.S. Geological Survey, Sioux Falls, South Dakota, 2004. 80

[78] William Easterly. *The tyranny of experts: Economists, dictators, and the forgotten rights of the poor*. Basic Books New York, 2014. 2

[79] Charles Ehlschlaeger. Using the at search algorithm to develop hydrologic models from digital elevation data. In *International Geographic Information Systems (IGIS) Symposium*, volume 89, pages 275–281, 1989. 112

[80] James A Espey and Molly Espey. Turning on the lights: a meta-analysis of residential electricity demand elasticities. *Journal of Agricultural and Applied Economics*, 36(1): 65–82, 2004. 83, 105

[81] Shu Fan and Rob J Hyndman. The price elasticity of electricity demand in south australia. *Energy Policy*, 39(6):3709–3719, 2011. 98

[82] S Fatichi, V Yu Ivanov, and E Caporali. Investigating interannual variability of precipitation at the global scale: Is there a connection with seasonality? *Journal of Climate*, 25(16):5512–5523, 2012. 34

[83] Massimo Filippini and Shonali Pachauri. Elasticities of electricity demand in urban indian households. *Energy policy*, 32(3):429–436, 2004. 83, 88, 90, 97, 164

[84] P. Fraenkel, O. Paish, V. Bokalders, A. Harvey, A. Brown, and R. Edwards. *Micro-hydro power: A guide for Development workers*. MDG Publishing, Water resource library, 1991. 83

[85] D.J. Fulford, P. Mosley, and A. Gill. Recommendations on the use of micro-hydro power in rural development. *Journal of international Development*, 12(7):975–983, 2000. 3, 4, 83, 105

[86] José M García-Ruiz, J Ignacio López-Moreno, Sergio M Vicente-Serrano, Teodoro Lasanta-Martínez, and Santiago Beguería. Mediterranean water resources in a global change scenario. *Earth-Science Reviews*, 105(3):121–139, 2011. 34

[87] B.B. Ghale, G.R. Shrestha, and R.J. DeLucia. Private micro-hydro power and associated investments in nepal: the barpak village case and broader issues. In *Natural resources forum*, volume 24, pages 273–284. Wiley Online Library, 2000. 3, 4, 5, 83, 104, 105, 108

[88] H.K. Ghimire. Harnessing of minis scale hydropower for rural electrification. *Hydro Nepal*, 1(2), 2008. 83

[89] AR Gilmour, R Thompson, and BR Cullis. Average information reml: an efficient algorithm for variance parameter estimation in linear mixed models. *Biometrics*, 51 (4):1440–1450, 1995. 67

[90] Arthur Gilmour, Brian Cullis, Sue Welham, Beverley Gogel, and Robin Thompson. An efficient computing strategy for prediction in mixed linear models. *Computational statistics & data analysis*, 44(4):571–586, 2004. 62

[91] Diana Glassman, Michele Wucker, Tanushree Isaacman, and Corinne Champilou. The water-energy nexus: Adding water to the energy agenda. *World Policy Institute*, 2011. 1

[92] Global Runoff Data Center. *Global Runoff Data Base*. Global Runoff Data Centre. Koblenz, Federal Institute of Hydrology (BfG), 2014. 80

[93] Michael F Goodchild. Citizens as sensors: the world of volunteered geography. *GeoJournal*, 69(4):211–221, 2007. 121

[94] P Goovaerts. Geostatistical approaches for incorporating elevation into the spatial interpolation of rainfall. *Journal of hydrology*, 228(1):113–129, 2000. 60

[95] Pierre Goovaerts. Geostatistics in soil science: state-of-the-art and perspectives. *Geoderma*, 89(1):1–45, 1999. 60

[96] JL Gordon. Hydropower cost estimates. *International water power & dam construction*, 35(11):30–37, 1983. 84

[97] L Gottschalk, I Krasovskaia, E Leblois, E Sauquet, et al. Mapping mean and variance of runoff in a river basin. *Hydrology and Earth System Sciences Discussions Discussions*, 3(2):299–333, 2006. 60, 61

[98] L. Gudmundsson, J.B. Bremnes, J.E. Haugen, and T. Engen Skaugen. Technical note: Downscaling RCM precipitation to the station scale using quantile mapping – a comparison of methods. *Hydrol. Earth Syst. Discuss.*, 9:6185–6201, 2012. doi: 10.5194/hessd-9-6185-2012. 9, 11, 18

[99] H.V. Gupta, G. Blöschl, J.J McDonnell, H.H.G Savenije, M. Sivapalan, A. Viglione, and T. Wagener. Outcomes of synthesis. In *Runoff Prediction in Ungauged Basins: Synthesis across Processes, Places and Scales*. Cambridge University Press, 2013. 79

[100] Anup Gurung, Om Prakash Gurung, and Sang Eun Oh. The potential of a renewable energy technology for rural electrification in nepal: A case study from tangting. *Renewable Energy*, 36(11):3203–3210, 2011. 84

[101] Omid Bozorg Haddad, Mahdi Moradi-Jalal, and Miguel A Mariño. Design–operation optimisation of run-of-river power plants. *Proceedings of the ICE-Water Management*, 164(9):463–475, 2011. 114

[102] G Brent Hall, Raymond Chipeniuk, Robert D Feick, Michael G Leahy, and Vivien Deparday. Community-based production of geographic information using open source software and web 2.0. *International journal of geographical information science*, 24(5): 761–781, 2010. 121

[103] W Michael Hanemann. Valuing the environment through contingent valuation. *The Journal of Economic Perspectives*, pages 19–43, 1994. 114

[104] Charles R Henderson. Best linear unbiased estimation and prediction under a selection model. *Biometrics*, pages 423–447, 1975. 67

[105] HKH-FRIEND. *Hindu Kush Himalayan - Flow Regimes from International Experimental and Network Data*. UNESCO International Hydrological Programme, 2004. 22, 48, 107

[106] George Hondroyiannis. Estimating residential demand for electricity in greece. *Energy Economics*, 26(3):319–334, 2004. 83, 105

[107] SMH Hosseini, Farshid Forouzbakhsh, and M Rahimpoor. Determination of the optimal installation capacity of small hydro-power plants through the use of technical, economic and reliability indices. *Energy Policy*, 33(15):1948–1956, 2005. 57

[108] Wen-Cheng Huang and Fu-Ti Yang. Streamflow estimation using kriging. *Water resources research*, 34(6):1599–1608, 1998. 60

[109] G.J. Huffman and D.T. Bolvin. Trmm and other data precipitation data set documentation. Technical report, Mesoscale Atmospheric Processes Laboratory, NASA Goddard Space Flight Center, 2013. 11

[110] G.J. Huffman, D.T. Bolvin, E.J. Nelkin, D.B. Wolff, R.F. Adler, G. Gu, Y. Hong, K.P. Bowman, and E.F. Stocker. The TRMM multisatellite precipitation analysis (TMPA): Quasi-global, multiyear, combined-sensor precipitation estimates at fine scales. *Journal of Hydrometeorology*, 8(1):38–55, 2007. doi: 10.1175/JHM560.1. 8, 107

[111] M. Huntermann. L'eau potable génératrice d'électricité, inventaire et étude du potentiel des usines éléectriques sur l'alimentation en eau potable en suisse. *DIANE*, 1994. 1

[112] IEA. World energy outlook. 2011. 1, 103

[113] R. Ihaka and R. Gentleman. R: A language for data analysis and graphics. *Journal of Computational and Graphical Statistics*, 5(3):299–314, 1996. doi: 10.1080/10618600.1996.10474713. 19, 69

[114] Atsushi Iimi. The impacts of metering and climate conditions on residential electricity demand: the case of albania. 2011. 83

[115] G. Imbens. Lecture notes. In *ARE2013: Applied Econometrics*, Agriculture and Resource Economics, University of California, Berkeley, 2006. 92

[116] W.W. Immerzeel, L.P.H. van Beek, and M.F.P. Bierkens. Climate change will affect the Asian water towers. *Science*, 328(5984):1382–1385, 2010. doi: 10.1126/science.1183188. 21

[117] M.N. Islam, S. Das, and H. Uyeda. Calibration of TRMM derived rainfall over Nepal during 1998–2007. *The Open Atmospheric Science Journal*, 4:12–23, 2010. doi: 10.2174/1874282301004010012. 8, 22

[118] Ignacio Rodriguez Iturbe and Amilcare Porporato. *Ecohydrology of Water-controlled Ecosystems: Soil Moisture and Plant Dynamics*. Cambridge University Press New York, 2004. 38

[119] IUCN, IWMI, Rasmar Convention Bureau and WRI World Resources Institute. *Watersheds of the World*. Washington, DC, 2003. 80

[120] Hanan G Jacoby. Access to markets and the benefits of rural roads. *The Economic Journal*, 110(465):713–737, 2000. 95

[121] O.D. Jimoh and P. Webster. The optimum order of a Markov chain model for daily rainfall in Nigeria. *Journal of Hydrology*, 185(1):45–69, 1996. doi: 10.1007/s00704-008-0051-3. 9, 12

[122] R.D. Joshi and V.B Amatya. Mini and micro hydropower development in the hindu kush-himalayan region-the nepal perspective. *ICIMOD*, 1996. 1, 4, 83, 108

[123] Chatchai Jothityangkoon and Murugesu Sivapalan. Temporal scales of rainfall–runoff processes and spatial scaling of flood peaks: space–time connection through catchment water balance. *Advances in water resources*, 24(9):1015–1036, 2001. 34

[124] A.A. Junejo. Manual for private and community-based mini-micro hydropower development in the hindukush himalayas. *ICIMOD*, 1997. 83, 111, 115

[125] A.A. Junejo, A. Karki, and R.K.L Karna. Manual for provate micro-hydropower - survey and layout design. *ICIMOD*, 1999. 83, 111

[126] S.R. Kansakar, D.M. Hannah, J. Gerrard, and G. Rees. Spatial pattern in the precipitation regime of Nepal. *International Journal of Climatology*, 24(13):1645–1659, 2004. doi: 10.1002/joc.1098. 8, 21, 22

[127] Timothy H Keitt, Roger Bivand, Edzer Pebesma, and Barry Rowlingson. rgdal: bindings for the geospatial data abstraction library. *R package version 0.7-1, URL http://CRAN. R-project. org/package= rgdal*, 2011. 106

[128] S. Khennas, A. Barnett, et al. Best practices for sustainable development of micro hydropower in developing countries - final synthesis report. *Department for International Development, UK.*, 2000. 1, 4, 5, 82, 95, 104, 105, 115

[129] James W Kirchner. Catchments as simple dynamical systems: Catchment characterization, rainfall-runoff modeling, and doing hydrology backward. *Water Resources Research*, 45(2), 2009. 47, 51

[130] BC Kusre, DC Baruah, PK Bordoloi, and SC Patra. Assessment of hydropower potential using GIS and hydrological modeling technique in Kopili River basin in Assam (India). *Applied Energy*, 87(1):298–309, 2010. 58, 111

[131] G Laaha, JO Skøien, and G Blöschl. Spatial prediction on river networks: comparison of top-kriging with regional regression. *Hydrological Processes*, 28(2):315–324, 2014. 61

[132] Gregor Laaha, Jon O Skøien, Franz Nobilis, and Günter Blöschl. Spatial prediction of stream temperatures using top-kriging with an external drift. *Environmental Modeling & Assessment*, 18(6):671–683, 2013. 61

[133] F Laio, A Porporato, L Ridolfi, and Ignacio Rodriguez-Iturbe. Plants in water-controlled ecosystems: Active role in hydrologic processes and response to water stress: Ii. probabilistic soil moisture dynamics. *Advances in Water Resources*, 24(7):707–723, 2001. 9, 11

[134] Shaun A Langley and Joseph P Messina. Embracing the open-source movement for managing spatial data: A case study of african trypanosomiasis in kenya. *Journal of map & geography libraries*, 7(1):87–113, 2011. 105

[135] D.G. Larentis, W. Collischonn, F. Olivera, and C.E.M. Tucci. Gis-based procedures for hydropower potential spotting. *Energy*, 2010. 58, 111

[136] RM Lark, BR Cullis, and SJ Welham. On spatial prediction of soil properties in the presence of a spatial trend: the empirical best linear unbiased predictor (e-blup) with reml. *European Journal of Soil Science*, 57(6):787–799, 2006. 62, 64, 67, 68

[137] R. Lee, J. Brizzee, S. Cherry, and D.G. Hall. Virtual Hydropower Prospecting: A Foundation for Water Energy Resource Planning and Development. *Journal of Map & Geography Libraries*, 4(2):336–347, 2008. ISSN 1542-0353. 58, 111

[138] GNU General Public License. version 3. *Free Software Foundation. http://www. gnu. org/copyleft/gpl. html. Last checked May*, 13:2010, 2007. 106

[139] Molly Lipscomb, Ahmed Mushfiq Mobarak, and Tania Barham. Development effects of electrification: Evidence from the topographic placement of hydropower plants in brazil. *Working Paper*, 2012. 84, 85

[140] John Loomis. What's to know about hypothetical bias in stated preference valuation studies? *Journal of Economic Surveys*, 25(2):363–370, 2011. 83, 125

[141] Brijesh Mainali and Semida Silveira. Renewable energy markets in rural electrification: country case nepal. *Energy for Sustainable Development*, 16(2):168–178, 2012. 3

[142] Fatemeh Masoudinia. Retscreen–small hydro project software. In *Communication Systems and Network Technologies (CSNT), 2013 International Conference on*, pages 858–861. IEEE, 2013. 105

[143] Arakaparampil M Mathai. *An introduction to geometrical probability: distributional aspects with applications*, volume 1. CRC Press, 1999. 157

[144] Daniel McFadden, Carlos Puig, and Daniel Kirschner. Determinants of the long-run demand for electricity. In *Proceedings of the American Statistical Association*, volume 1, pages 109–19. Business and Economics Section, 1977. 84

[145] M.A. McNeil and V.E. Letschert. Forecasting electricity demand in developing countries: A study of household income and appliance ownership. *European Council for an Energy Efficient Economy-2005 Summer Study. Mandelieu, France. LBNL-58283*, 2005. 85, 90

[146] V Meier, Jean-Max Baumer, et al. Harnessing water power on a small scale; local experience with micro-hydro technology. In *Harnessing water power on a small scale; local experience with micro-hydro technology*. Skat, 1982. 108

[147] Ralf Merz and Günter Blöschl. Flood frequency regionalisation - spatial proximity vs. catchment attributes. *Journal of Hydrology*, 302(1):283–306, 2005. 70

[148] M. Mikhail and R. Yoder. Multiple use water service implementation in nepal and india - experience and lessons for scale up. *IDE, CPWF and IWMI*, 2008. 1

[149] Lewis F Moody. Friction factors for pipe flow. *Trans. Asme*, 66(8):671–684, 1944. 167

[150] Rafael Moreno-Sanchez, Geoffrey Anderson, Jesus Cruz, and Mary Hayden. The potential for the use of open source software and open specifications in creating web-based cross-border health spatial information systems. *International Journal of Geographical Information Science*, 21(10):1135–1163, 2007. 105

[151] W. Mostert. Scaling-up micro-hydro, lessons from nepal and a few notes on solar home systems. *Village Power*, 1998. 4, 82, 104

[152] Marc F Müller and Sally E Thompson. Bias adjustment of satellite rainfall data through stochastic modeling: Methods development and application to nepal. *Advances in Water Resources*, 60:121–134, 2013. 6, 69, 71, 86

[153] Marc F Müller, David N Dralle, and Sally E Thompson. Analytical model for flow duration curves in seasonally dry climates. *Water Resources Research*, 50(7):5510–5531, 2014. 6, 108

[154] MF Müller and SE Thompson. A topological restricted maximum likelihood (topreml) approach to regionalize trended runoff signatures in stream networks. *Hydrology and Earth System Sciences Discussions*, 12(1):1355–1396, 2015. 6

[155] R Muneepeerakul, A Rinaldo, and I Rodriguez-Iturbe. Effects of river flow scaling properties on riparian width and vegetation biomass. *Water resources research*, 43 (12), 2007. 57

[156] H MYERS Myers, Raymond H. Classical and modern regression with applications. *PWS-KENT Publishing Compagny*, 1990. 59

[157] F. Naronha. Open source software opens new windows to third world. *Linux Journal*, 2002. 106

[158] F. Naronha. Developing countries gain from free/open source softwares. *Linux Journal*, 2003. 106

[159] Kapil Narula, Yu Nagai, and Shonali Pachauri. The role of decentralized distributed generation in achieving universal rural electrification in south asia by 2030. *Energy Policy*, 47:345–357, 2012. 103

[160] NASA. http://mirador.gsfc.nasa.gov, November 2012. 22

[161] NASA EOSDIS Global Hydrology Resource Center (GHRC) DAAC. *LIS Global lightning Image*. Huntsville, AL, 2011. 80

[162] NASA Socioeconomic Data and Applications Center (SEDAC). *National Aggregates of Geospatial Data Collection: Population, Landscape, And Climate Estimates, Version 3 (PLACE III)*. 2012. 34, 86

[163] RJ Nathan and TA McMahon. Evaluation of automated techniques for base flow and recession analyses. *Water Resources Research*, 26(7):1465–1473, 1990. 44, 46, 54

[164] Nepal Micro Hydropower Development Association, 2015. URL http://www.microhydro.org.np. 3

[165] M. Neuhaus. Salleri chialsa project. *Water Power and Dam Construction*, September 1994. 3

[166] Central Bureau of Statistics. *Nepal Living Standards Survey (I,II,III)*. Government of Nepal, 2012. 87, 89, 164

[167] R.A. Olea. Optimal contour mapping using universal kriging. *Journal of Geophysical Research*, 79(5):695–702, 1974. doi: 10.1029/JB079i005p00695. 60

[168] S. Pachauri. An analysis of cross-sectional variations in total household energy requirements in india using micro survey data. *Energy policy*, 32(15):1723–1735, 2004. 85, 90

[169] Oliver Paish. Micro-hydropower: status and prospects. *Proceedings of the Institution of Mechanical Engineers, Part A: Journal of Power and Energy*, 216(1):31–40, 2002. 93, 103

[170] H Desmond Patterson and Robin Thompson. Recovery of inter-block information when block sizes are unequal. *Biometrika*, 58(3):545–554, 1971. 62

[171] Edzer J. Pebesma. Multivariable geostatistics in S: the gstat package. *Computers and Geosciences*, 30:683–691, 2004. doi: 10.1016/j.cageo.2004.03.012. 16, 71

[172] Andre Plourde and David Ryan. On the use of double-log forms in energy demand analysis. *The Energy Journal*, pages 105–113, 1985. 88, 97

[173] S. Pokharel. An econometric analysis of energy consumption in nepal. *Energy policy*, 35(1):350–361, 2007. 82, 85

[174] Amilcare Porporato, Edoardo Daly, and Ignacio Rodriguez-Iturbe. Soil water balance and ecosystem response to climate change. *The American Naturalist*, 164(5):625–632, 2004. 47

[175] Satya Prakash, C Mahesh, and RM Gairola. Comparison of TRMM Multi-satellite Precipitation Analysis (TMPA)-3B43 version 6 and 7 products with rain gauge data from ocean buoys. *Remote Sensing Letters*, 4(7):677–685, 2013. doi: 10.1080/2150704X.2013.783248. 22

[176] A Pugliese, A Castellarin, and A Brath. Geostatistical prediction of flow-duration curves. *Hydrology and Earth System Sciences Discussions*, 10(11):13053–13091, 2013. 46

[177] Dario Pumo, Leonardo Valerio Noto, and Francesco Viola. Ecohydrological modelling of flow duration curve in mediterranean river basins. *Advances in Water Resources*, 52:314–327, 2013. 56

[178] Petras Punys, Antanas Dumbrauskas, Algis Kvaraciejus, and Gitana Vyciene. Tools for small hydropower plant resource planning and development: A review of technology and applications. *Energies*, 4(9):1258–1277, 2011. 104

[179] Peter C Reiss and Matthew W White. Household electricity demand, revisited. *The Review of Economic Studies*, 72(3):853–883, 2005. 84, 88, 98

[180] Renewable Energy for Rural Livelihood Programme. Detailed project reports. Technical report, Alternative Energy Promotion Centre (AEPC), Governement of Nepal, http://www.rerl.org.np/publication/cat.php?c=10, 2009. 91, 96, 111, 167

[181] M.E. Renwick. Multiple use water services for the poor: Assessing the state of knowledge. *Winrock International*, 2007. 1

[182] Maura Rianna, Andreas Efstratiadis, Fabio Russo, Francesco Napolitano, and Demetris Koutsoyiannis. A stochastic index method for calculating annual flow duration curves in intermittent rivers. *Irrigation and Drainage*, 62(S2):41–49, 2013. 42, 56

[183] P.J Ribeiro and P.J Diggle. geoR: a package for geostatistical analysis. *R-NEWS*, 1(2): 14–18, June 2001. URL http://CRAN.R-project.org/doc/Rnews/. ISSN 1609-3631. 16

[184] C.W. Richardson. Stochastic simulation of daily precipitation, temperature, and solar radiation. *Water Resources Research*, 17(1):182–190, 1981. doi: 10.1029/WR017i001p00182. 9, 12

[185] Brian Richter, Jeffrey Baumgartner, Robert Wigington, and David Braun. How much water does a river need? *Freshwater biology*, 37(1):231–249, 1997. 57

[186] Brian D Richter, Ruth Mathews, David L Harrison, and Robert Wigington. Ecologically sustainable water management: managing river flows for ecological integrity. *Ecological applications*, 13(1):206–224, 2003. 57

[187] Brian D Ripley. *Spatial statistics*, volume 575. John Wiley & Sons, 2005. 118, 168

[188] Ignacio Rodriguez-Iturbe and José M Mejía. On the transformation of point rainfall to areal rainfall. *Water Resources Research*, 10(4):729–735, 1974. doi: 10.1029/WR010i004p00729. 15, 149, 150

[189] Javier Rosa, P Achintya Madduri, and Daniel Soto. Efficient microgrid management system for electricity distribution in emerging regions. In *Global Humanitarian Technology Conference (GHTC), 2012 IEEE*, pages 23–26. IEEE, 2012. 83

[190] Juan Pablo Rud. Electricity provision and industrial development: Evidence from india. *Journal of Development Economics*, 97(2):352–367, 2012. 84, 85

[191] Jos M Samuel, Murugesu Sivapalan, and Iain Struthers. Diagnostic analysis of water balance variability: A comparative modeling study of catchments in perth, newcastle, and darwin, australia. *Water resources research*, 44(6), 2008. 34

[192] A Santolin, G Cavazzini, G Pavesi, G Ardizzon, and A Rossetti. Techno-economical method for the capacity sizing of a small hydropower plant. *Energy Conversion and Management*, 52(7):2533–2541, 2011. 57

[193] Eric Sauquet. Mapping mean annual river discharges: geostatistical developments for incorporating river network dependencies. *Journal of Hydrology*, 331(1):300–314, 2006. 60, 61

[194] Eric Sauquet, Lars Gottschalk, and Etienne Leblois. Mapping average annual runoff: a hierarchical approach applying a stochastic interpolation scheme. *Hydrological sciences journal*, 45(6):799–815, 2000. 60, 61

[195] Bettina Schaefli, Andrea Rinaldo, and Gianluca Botter. Analytic probability distributions for snow-dominated streamflow. *Water Resources Research*, 49(5):2701–2713, 2013. 35, 37, 81, 124

[196] J.L. Schenker. Open source software gets boost at un. *International Herald Tribune*, 2003. 106

[197] Peter Schmidt. Three-stage least squares with different instruments for different equations. *Journal of econometrics*, 43(3):389–394, 1990. 92, 166

[198] Andrew Scott and Prachi Seth. The political economy of electricity distribution in developing countries. *Journal of policy and Governance*, 2013. 98

[199] James Kincheon Searcy. *Flow-duration curves*. US Government Printing Office, 1959. 34

[200] Jasjeet S Sekhon. Multivariate and propensity score matching software with automated balance optimization: the matching package for r. *Journal of Statistical Software*, 42 (7):1–52, 2011. 87

[201] J. Shao and D. Tu. *The Jackknife and Bootstrap*. Springer-Verlag, New York, 1995. 19, 24

[202] G Shi, X Zheng, and F Song. Estimating elasticity for residential electricity demand in china. *The Scientific World Journal*, 2012, 2012. 83

[203] M.S. Shrestha. *Bias-Adjustment of Satellite-Based Rainfall Estimates over the Central Himalayas of Nepal for Flood Prediction*. PhD thesis, Kyoto University, 2011. 8

[204] Julian I Silk and Frederick L Joutz. Short and long-run elasticities in us residential electricity demand: a co-integration approach. *Energy Economics*, 19(4):493–513, 1997. 88, 164

[205] SK Singal and RP Saini. Cost analysis of low-head dam-toe small hydropower plants based on number of generating units. *Energy for Sustainable development*, 12(3):55–60, 2008. 84, 85, 89

[206] SK Singal, RP Saini, and CS Raghuvanshi. Analysis for cost estimation of low head run-of-river small hydropower schemes. *Energy for sustainable Development*, 14(2): 117–126, 2010. 84, 85, 89

[207] M Sivapalan, M Konar, V Srinivasan, A Chhatre, A Wutich, CA Scott, JL Wescoat, and I Rodríguez-Iturbe. Socio-hydrology: Use-inspired water sustainability science for the anthropocene. *Earth's Future*, 2(4):225–230, 2014. 59

[208] JO Skøien, R Merz, and G Blöschl. Top-kriging–geostatistics on stream networks. *Hydrology & Earth System Sciences*, 10(2), 2006. 60, 61, 77, 78

[209] Jon Olav Skøien, G Blöschl, G Laaha, E Pebesma, J Parajka, and A Viglione. rtop: An r package for interpolation of data with a variable spatial support, with an example from river networks. *Computers & Geosciences*, 67:180–190, 2014. 69, 70

[210] Dean F Smith, Albin J Gasiewski, Darren L Jackson, and Gary A Wick. Spatial scales of tropical precipitation inferred from trmm microwave imager data. *Geoscience and Remote Sensing, IEEE Transactions on*, 43(7):1542–1551, 2005. 80

[211] KER Soetaert, Thomas Petzoldt, and R Woodrow Setzer. Solving differential equations in r: package desolve. *Journal of Statistical Software*, 33, 2010. 47

[212] Didier Sornette. *Critical phenomena in natural sciences: chaos, fractals, selforganization and disorder: concepts and tools*. Springer Science & Business Media, 2004. 41

[213] R. Srikanthan and T.A. McMahon. Stochastic generation of annual, monthly and daily climate data: A review. *Hydrology and Earth System Sciences*, 5(4):653–670, 2001. doi: 10.5194/hess-5-653-2001. 9, 12

[214] V Srinivasan, S Thompson, K Madhyastha, G Penny, K Jeremiah, and S Lele. Why is the arkavathy river drying? a multiple hypothesis approach in a data scarce region. *Hydrology and Earth System Sciences Discussions*, 12(1):25–66, 2015. 59

[215] Stefan Steiniger and Andrew JS Hunter. The 2012 free and open source gis software map–a guide to facilitate research, development, and adoption. *Computers, Environment and Urban Systems*, 39:136–150, 2013. 105

[216] M Stoelzle, K Stahl, and M Weiler. Are streamflow recession characteristics really characteristic? *Hydrology and Earth System Sciences*, 17(2):817–828, 2013. 44

[217] Erik Stokstad. Scarcity of rain, stream gages threatens forecasts. *Science*, 285(5431): 1199–1200, 1999. 59

[218] LM Tallaksen. A review of baseflow recession analysis. *Journal of hydrology*, 165(1): 349–370, 1995. 44

[219] GRASS Development Team. *Geographic Resources Analysis Support System (GRASS) Software*. Open Source Geospatial Foundation, 2012. 106

[220] The World Bank. *Electric power transmission and distribution losses*. Retrieved from http://data.worldbank.org/indicator/EG.ELC.LOSS.ZS, 2015. 82

[221] Sally Thompson, Simon Levin, and Ignacio Rodriguez-Iturbe. Linking plant disease risk and precipitation drivers: a dynamical systems framework. *The American Naturalist*, 181(1):E1–E16, 2013. 57

[222] SE Thompson, M Sivapalan, CJ Harman, V Srinivasan, MR Hipsey, P Reed, A Montanari, and G Blöschl. Developing predictive insight into changing water systems: use-inspired hydrologic science for the anthropocene. *Hydrology and Earth System Sciences*, 17(12):5013–5039, 2013. 57

[223] UNEP-WCMC. *Mountains and Forests in Mountains*. Cambridge, 2000. 80

[224] United States Geological Survey. *National Water Information System*. 2013. 48

[225] United States National Aeronautics and Space Administration (NASA) and Ministry of Economy, Trade, and Industry (METI) of Japan. ASTER Global Digital Elevation Model v2. `http://asterweb.jpl.nasa.gov/gdem.asp`, October 2011. 20, 86, 107

[226] S.L. Vaidya. Cost and revenue structures for micro-hydro projects in nepal. Technical report, Alternative Energy Promotion Center, Nepal. 83, 85

[227] B VanKoppen, S. Smits, P. Moriarty, F. Penning de Vries, Mikhail M., and B Boelee. Climbing the water ladder: Multiple-use water services for poverty reduction. *IRC International Water and Sanitation Center and International Water Management Institute*, 2009. 1

[228] W.N. Venables and B.D. Ripley. *Modern applied statistics with S*. Springer-Verlag, New York, 2002. 24

[229] Jay M Ver Hoef and Erin E Peterson. A moving average approach for spatial statistical models of stream networks. *Journal of the American Statistical Association*, 105(489), 2010. 61, 62

[230] Jay M Ver Hoef, Erin Peterson, and David Theobald. Spatial statistical models that use flow and stream distance. *Environmental and Ecological statistics*, 13(4):449–464, 2006. 61

[231] A Viglione, J Parajka, M Rogger, JL Salinas, G Laaha, M Sivapalan, and G Blöschl. Comparative assessment of predictions in ungauged basins–part 3: Runoff signatures in austria. *Hydrology and Earth System Sciences*, 17(6):2263–2279, 2013. 61

[232] Richard M Vogel and Neil M Fennessey. Flow-duration curves. i: New interpretation and confidence intervals. *Journal of Water Resources Planning and Management*, 120 (4):485–504, 1994. 35, 42

[233] Richard M Vogel and Neil M Fennessey. Flow duration curves ii: A review of applications in water resources planning1, 1995. 35

[234] E. Ward, W. Buytaert, L. Peaver, and H. Wheater. Evaluation of precipitation products over complex mountainous terrain: A water resources perspective. *Advances in Water Resources*, 34(10):1222–1231, 2011. doi: 10.1016/j.advwatres.2011.05.007. 8

[235] Water Information Provision Section, Perth Western Australia. *Hydstra database*. 2013. 48

[236] WHO and UNICEF. Progress on sanitation and drinking water. *JMP MDG update report*, 2010. 1

[237] D.S. Wilks. Multisite generalization of a daily stochastic precipitation generation model. *Journal of Hydrology*, 210(1):178–191, 1998. doi: 10.1016/S0022-1694(98) 00186-3. 11

[238] H Wittenberg and M Sivapalan. Watershed groundwater balance estimation using streamflow recession analysis and baseflow separation. *Journal of Hydrology*, 219(1): 20–33, 1999. 69

[239] C. Wolfram, O. Shelef, and P.J. Gertler. How will energy demand develop in the developing world? Technical report, National Bureau of Economic Research, 2012. 85, 90

[240] JM Woodridge. Econometric analysis of cross sectional data and panel data, 2002. 92

[241] World Bank. Renewable energy powers rural nepal into the future. URL http://www.worldbank.org/en/news/feature/2014/02/05/renewable-energy-powers-rural-nepal-into-the-future. 3

[242] SJ Wright and J Nocedal. *Numerical optimization*, volume 2. Springer New York, 1999. 69, 159

[243] Tingbao Xu, Barry Croke, and Michael F Hutchinson. Identification of spatial and temporal patterns of australian daily rainfall under a changing climate. In D.P. Ames, N.W.T. Quinn, and A.E. Rizzoli, editors, *Proceedings of the 7th International Congress on Environmental Modelling and Software*, 2014. 80

[244] M.K. Yamamoto, K. Ueno, and K. Nakamura. Comparison of satellite precipitation products with rain gauge data for the Khumb region, Nepal Himalayas. *Journal of the Meteorological Society of Japan*, 89(6):597–610, 2011. doi: 10.2151/jmsj.2011-601. 22

[245] Soni Yatheendradas, Thorsten Wagener, Hoshin Gupta, Carl Unkrich, David Goodrich, Mike Schaffner, and Anne Stewart. Understanding uncertainty in distributed flash flood forecasting for semiarid regions. *Water Resources Research*, 44(5), 2008. doi: 10.1029/2007WR005940. 8

[246] Sheng Ye, M Yaeger, E Coopersmith, L Cheng, and M Sivapalan. Exploring the physical controls of regional patterns of flow duration curves–part 2: Role of seasonality, the regime curve, and associated process controls. *Hydrology and Earth System Sciences*, 16(11):4447–4465, 2012. 35

[247] A. Yee. Microhydro drives change in rural nepal. *The New Your Times*, 6(22), 2012. 3

[248] Choong-Sung Yi, Jin-Hee Lee, and Myung-Pil Shim. Site location analysis for small hydropower using geo-spatial information system. *Renewable Energy*, 35(4):852–861, 2010. 58, 111

[249] Z.Y. Yin, X. Zhang, X. Liu, M. Colella, and X. Chen. An assessment of the biases of satellite rainfall estimates over the Tibetan Plateau and correction methods based on topographic analysis. *Journal of Hydrometeorology*, 9(3):301–326, 2008. doi: 10.1175/2007JHM903.1. 8, 9, 22

[250] Arnold Zellner and Henri Theil. Three-stage least squares: simultaneous estimation of simultaneous equations. *Econometrica: Journal of the Econometric Society*, pages 54–78, 1962. 92

Appendix A

Supplementary Material to Chapter 2

A.1 From stochastic model parameters to evaluation metrics

The output of our stochastic model are the 10 parameters described in Table 2.2. Combining these parameters, one can obtain seasonal metrics such as the unconditional expectation and variance of daily rainfall ($\mathsf{E}\left[X^{(i)}\right]$, $\mathsf{Var}\left(X^{(i)}\right)$) the expected length of wet and dry spells ($L_w^{(i)}, L_d^{(i)}$) and the probability of rainfall occurring on any given day ($P^{(i)}$). These relationships are listed in Equation (A.1)-Equation (A.5):

$$L_w^{(i)} = \frac{1}{1 - P_{11}^{(i)}} \tag{A.1}$$

$$L_d^{(i)} = \frac{1}{P_{01}^{(i)}} \tag{A.2}$$

$$P^{(i)} = \frac{P_{01}^{i}}{1 + P_{01}^{(i)} - P_{11}^{(i)}} \tag{A.3}$$

$$\mathsf{E}\left[X^{(i)}\right] = P^{(i)}\frac{\mathsf{GS}}{\mathsf{GR}} \tag{A.4}$$

$$\mathsf{Var}\left(X^{(i)}\right) = P^{(i)}\frac{\mathsf{GS}}{\mathsf{GR}^2} + P^{(i)}(1 - P^{(i)})\frac{\mathsf{GS}}{\mathsf{GR}} \tag{A.5}$$

where the subscript i indicates either the wet (w) or dry (d) season. By weighing seasonal metrics by the duration of the corresponding season I get the annual metrics (L_w, P, $\mathsf{E}\left[X\right]$ and $\mathsf{Var}\left(X\right)$):

$$P = L_{\mathsf{Rn}} P^{(w)} + (1 - L_{\mathsf{Rn}}) P^{(d)} \tag{A.6}$$

$$L_w = \left(1 - \frac{L_{\mathsf{Rn}} P^{(w)} P_{11}^{(w)} + (1 - L_{\mathsf{Rn}}) P^{(d)} P_{11}^{(d)}}{P}\right)^{-1} \tag{A.7}$$

$$\mathsf{E}\left[X\right] = L_{\mathsf{Rn}} P^{(w)} \frac{\mathsf{GS}^{(w)}}{\mathsf{GR}^{(w)}} + (1 - L_{\mathsf{Rn}}) P^{(d)} \frac{\mathsf{GS}^{(d)}}{\mathsf{GR}^{(d)}} \tag{A.8}$$

$$\begin{aligned}\mathsf{Var}\left(X\right) = {} & L_{\mathsf{Rn}} \left(\frac{\mathsf{GS}^{(w)}}{\mathsf{GR}^{(w)}}\right)^2 + (1 - L_{\mathsf{Rn}}) \left(\frac{\mathsf{GS}^{(d)}}{\mathsf{GR}^{(d)}}\right)^2 - \left(L_{\mathsf{Rn}} \frac{\mathsf{GS}^{(w)}}{\mathsf{GR}^{(w)}} + (1 - L_{\mathsf{Rn}}) \frac{\mathsf{GS}^{(d)}}{\mathsf{GR}^{(d)}}\right)^2 \\ & + L_{\mathsf{Rn}} \frac{\mathsf{GS}^{(w)}}{\mathsf{GR}_{(w)}^2} + (1 - L_{\mathsf{Rn}}) \frac{\mathsf{GS}^{(d)}}{\mathsf{GR}_{(d)}^2} \end{aligned} \tag{A.9}$$

$$\tag{A.10}$$

with $L_{\mathsf{Rn}} = (\mathsf{RnStp} - \mathsf{RnStr})/365$, the fraction of the year occupied by the rainy season. Finally, yearly rainfall and the average number of rainy days per year can easily be obtained by multiplying $\mathsf{E}\left[X\right]$ and P by 365 respectively.

A.2 Aggregation of rainfall occurrence probabilities

A pixel is in a rainy state on a given day if it rains at *any* of its gauges during that day, which precludes an area weighting approach from being applied to aggregate rain occurrence parameters. Indeed, let a pixel contain two gauges with equal weights and rainfall probabilities of 0.1 and 1 respectively: because it rains every day at one of the gauges, rainfall probability at the pixel level will be 1, which is not the average of the probabilities at the gauges.

Assuming a pixel contains N_p gauges with rainfall probabilities P_i , the following bounds apply:

$$\max P_i \leq P_{\mathsf{pix}} \leq \min \left\{ \sum_{i=1}^{N_p} P_i, 1 \right\} \tag{A.11}$$

P_{pix} reaches the lower bound if the correlation between rain occurrence is positive and maximal, i.e. a dry day at the gauge with highest P always corresponds to a dry day for the pixel. The higher bond is reached if the correlation is negative with a maximal absolute value, i.e. it almost always rains on at least one of the gauges.

In order to satisfy the two degrees of freedom offered by the two Markov transition probabilities (P_{01} and P_{11}) considered as SMPs, a the aggregation of a second metric (other than P_i) must be considered. The pixel aggregated value of $P_{\mathsf{pix}} \cdot P_{11,\mathsf{pix}}$, the ratio of a wet-to-wet transitions, is bounded by

$$\max\left(P_i \cdot P_{11,i}\right) \leq P_{\mathsf{pix}} \cdot P_{11,\mathsf{pix}}$$

because such a transition occurring at a gauge is a *sufficient* condition for it to be aggregated at the pixel level. Similarly,$P_{\mathsf{pix}} \cdot P_{10,\mathsf{pix}}$, the ratio of a wet-to-dry transitions is bounded by

$$\sum_{i=1}^{N_p} P_i \cdot P_{10,i} \geq P_{\mathsf{pix}} \cdot P_{10,\mathsf{pix}}$$

because such a transition occurring at the gauge level is a *necessary* condition for it to be aggregated at the pixel level. Finally, both transition ratios are bounded by the maximum probability of rainfall according to the inequality in Equation (A.11). Therefore, with $P_{11} = 1 - P_{10}$, the bounds on $P_{11,\mathsf{pix}}$ can be written as:

$$\max\left\{\frac{\max\left(P_i \cdot P_{11,i}\right)}{P_{\mathsf{pix}}}; 1 - \frac{\sum_{i=1}^{N_p} P_i \cdot (1 - P_{11,i})}{P_{\mathsf{pix}}}\right\} \leq P_{11,\mathsf{pix}} \leq \min\left\{\frac{\sum\limits_{i=1}^{N_p} P_i}{P_{\mathsf{pix}}}, 1\right\} \tag{A.12}$$

Within these bounds, rainfall probability and the ratio of a wet-to-wet transitions increase with the pixel size and the number of gauges within the pixel. The actual value of these metrics depends on the spatial auto-correlation of rainfall occurrences within the pixels. If rain occurrence is highly spatially auto-correlated, which is likely in pixels smaller than the spatial scale of typical rain events, I can approximate:

$$P_{\mathsf{pix}} \approx \max P_i \tag{A.13}$$

$$P_{11,\mathsf{pix}} \approx \max\left\{\frac{\max\left(P_i \cdot P_{11,i}\right)}{P_{\mathsf{pix}}}; 1 - \frac{\sum_{i=1}^{N_p} P_i \cdot (1 - P_{11,i})}{P_{\mathsf{pix}}}\right\} \tag{A.14}$$

With spatial autocorrelation ranges of approximately 3 (dry season) to 4 (wet season) times the pixel size of 27.7km considered in Nepal (Table 2.4), a Monte Carlo analysis showed that these approximations lead to an average underestimation of less then 2% for both metrics for up to five gauges per pixel. This error increases with the number of gauges and decreases with the range of spatial autocorrelation.

A.3 Aggregation of conditional rainfall depth distribution

Consider a square pixel of side d with N_p gauges, each covering a Thiessen Polygon of size a_i, where the weights a_i are normalized such that $\sum_{i=1}^{N_p} a_i = 1$. For each gauge i, I have access to daily precipitation data X_i, as well as the statistics $\mathsf{E}\left[X_i \mid \mathsf{wet}\right]$ and $\mathsf{Var}\left(X_i \mid \mathsf{wet}\right)$, measuring the mean and the variance of local rainfall on a rainy day respectively. Assuming

that the precipitation depth on wet days follows a gamma distribution, these statistics can be straightforwardly related to the shape (GS) and rate (GR) of that distribution:

$$\mathsf{E}\left[X \mid \mathsf{wet}\right] = \frac{GS}{GR}$$

$$\mathsf{Var}\left(X \mid \mathsf{wet}\right) = \frac{GS}{GR^2}$$

I wish to estimate $\mathsf{E}\left[X_{\mathsf{pix}} \mid \mathsf{wet}\right]$ and $\mathsf{Var}\left(X_{\mathsf{pix}} \mid \mathsf{wet}\right)$, the mean and variance of the areal rainfall on wet days aggregated at the pixel level, which will lead us to GS_{pix} and GR_{pix} the aggregated parameters of our stochastic model.

As a first step, I determine the local rainfall at a random point of the pixel X_{pt} according a two-step data generating process as follows:

(i) At the outset, before any measurements are made, a point of the pixel is chosen uniformly at random. As a result, the area weights a_i measure the probability that this point is located in Thiessen Polygon i.

(ii) Subsequently, I assume that local rainfall across the entire Thiessen Polygon is constant and measured by gauge i.

As a result, I can determine the expected unconditional rainfall at a random point of the pixel using the law of iterated expectation,

$$\mathsf{E}\left[X_{\mathsf{pt}}\right] = \mathsf{E}\left[\mathsf{E}\left[X_{\mathsf{pt}} \mid i\right]\right] = \sum_{i=1}^{N_p} a_i \mathsf{E}\left[X_i\right].$$

Here, $\mathsf{E}\left[X_{\mathsf{pt}} \mid i\right]$ denotes the expected rainfall conditional on the random point being in Polygon i, in which case X_{pt} is equal to X_i by assumption. Knowing that the mean value over the pixel area of all possible realizations of the point process X_{pt} results in an areal rainfall process with an identical expectation [188] (i.e. $\mathsf{E}\left[X_{\mathsf{pix}}\right] = \mathsf{E}\left[X_{\mathsf{pt}}\right]$), I can calculate the mean areal rainfall on a rainy day by applying the law of iterated expectations both at the pixel level and for each individual gauge,

$$\mathsf{E}\left[X_{\mathsf{pix}} \mid \mathsf{wet}\right] = \frac{1}{P_{\mathsf{pix}}} \cdot \sum_{i=1}^{N_p} a_i P_i \mathsf{E}\left[X_i \mid \mathsf{wet}\right]. \tag{A.15}$$

with P_i and P_{pix} the probability or rainfall at the gauge i and at the pixel level respectively and represent the expectations of the binomial stochastic processes defining rainfall occurrence at these points.

For the local variance, the same data generating process implies, by the law of total variance,

$$\begin{aligned}\mathsf{Var}\left(X_{\mathsf{pt}}\right) &= \mathsf{E}\left[\mathsf{Var}\left(X_{\mathsf{pt}} \mid i\right)\right] + \mathsf{Var}\left(\mathsf{E}\left[X_{\mathsf{pt}} \mid i\right]\right) \\ &= \mathsf{E}\left[\mathsf{Var}\left(X_i\right)\right] + \mathsf{E}\left[\mathsf{E}\left[X_i\right]^2\right] - \mathsf{E}\left[\mathsf{E}\left[X_i\right]\right]^2 \\ &= \sum_{i=1}^{N_p} a_i \mathsf{Var}\left(X_i\right) + \sum_{i=1}^{N_p} a_i \mathsf{E}\left[X_i\right]^2 - \left(\sum_{i=1}^{N_p} a_i \mathsf{E}\left[X_i\right]\right)^2\end{aligned}$$

Again, I condition on the polygon i and assume that precipitation is homogenous within each Thiessen polygon (i.e. $X_{\mathsf{pt}} = X_i$). From Equation (A.15), I get:

$$\mathsf{Var}\left(X_{\mathsf{pt}}\right) = \sum_{i=1}^{N_p} a_i \mathsf{Var}\left(X_i\right) + \sum_{i=1}^{N_p} a_i \mathsf{E}\left[X_i\right]^2 - P_{\mathsf{pix}}^2 \mathsf{E}\left[X_{\mathsf{pix}} \mid \mathsf{wet}\right]^2 \tag{A.16}$$

In the next step, I condition on rainfall probability, applying the law of total variance and taking rainfall occurrence as a binomial random variable:

$$\begin{aligned}\mathsf{Var}\left(X_i\right) &= \mathsf{Var}\left(\mathsf{E}\left[X_i \mid \mathsf{wet}\right]\right) + \mathsf{E}\left[\mathsf{Var}\left(X_i \mid \mathsf{wet}\right)\right] \\ &= P_i(1-P_i)\mathsf{E}\left[X_i \mid \mathsf{wet}\right] + P_i \mathsf{Var}\left(X_i \mid \mathsf{wet}\right)\end{aligned} \tag{A.17}$$

Substituting Equation (A.17) inEquation (A.16) I have:

$$\begin{aligned}&(1-P_{\mathsf{pix}})P_{\mathsf{pix}}\mathsf{E}\left[X_{\mathsf{pix}} \mid \mathsf{wet}\right] + P_{\mathsf{pix}}\mathsf{Var}\left(X_{\mathsf{pt}} \mid \mathsf{wet}\right) = \\ &= P_i(1-P_i)\mathsf{E}\left[X_i \mid \mathsf{wet}\right] + P_i\mathsf{Var}\left(X_i \mid \mathsf{wet}\right) + \sum_{i=1}^{N_p} a_i \mathsf{E}\left[X_i\right]^2 - P_{\mathsf{pix}}^2 \mathsf{E}\left[X_{\mathsf{pix}} \mid \mathsf{wet}\right]^2\end{aligned}$$

Using Equation (A.15) to express $P_{\mathsf{pix}}\mathsf{E}\left[X_{\mathsf{pix}} \mid \mathsf{wet}\right]$ and rearranging, I get the expression for the point variance:

$$\begin{aligned}&\mathsf{Var}\left(X_{\mathsf{pt}} \mid \mathsf{wet}\right) = \\ &\quad = \sum_{i=1}^{N_p} \frac{a_i P_i}{P_{\mathsf{pix}}}\left[\mathsf{Var}\left(X_i \mid \mathsf{wet}\right) + P_i^2 \mathsf{E}\left[X_i \mid \mathsf{wet}\right]^2 + (P_{\mathsf{pix}} - P_i)\mathsf{E}\left[X_i \mid \mathsf{wet}\right]\right] \\ &\quad - P_{\mathsf{pix}}\mathsf{E}\left[X_{\mathsf{pix}} \mid \mathsf{wet}\right]^2\end{aligned} \tag{A.18}$$

Finally, following Rodriguez-Iturbe and Mejía [188], I can infer the variance of area rainfall X_{pix} from that of the point rainfall process X_{pt} by correcting it with a factor

$$C(d) = \int^{\sqrt{2}d} r(\nu) f(\nu) \mathrm{d}\nu \leq 1,$$

where $r(\nu)$ is the spatial correlation function and $f(\nu)$ the distribution of distances between two points chosen at random in the pixel. In other words, $\mathsf{Var}\left(X_{\mathsf{pix}}\right) = C(d) \cdot \mathsf{Var}\left(X_{\mathsf{pt}}\right)$,

implying that point rainfall typically overestimates the variance of area rainfall because the area averaged intensity of local rainfall events are dampened by the absence of rain in parts of the pixels that do not fall in the current extent of the storm. It directly follows that the attenuation factor $C(d)$ is increasing in pixel size d and decreasing in spatial autocorrelation range. For TRMM pixels in Nepal, where pixel size and spatial auto-correlation are spatially homogenous, I have estimated $C(27.7\mathsf{km})$ at 0.75 in the monsoon and 0.86 in the dry season, using a correlogram estimated based on the spatial distribution of rainfall occurrences at gauges on 2,000 randomly drawn days. Therefore, I can express the conditional variance of areal rainfall at the pixel level as a function of the moments of conditional rainfall measured at the gauges:

$$\begin{aligned}\mathsf{Var}\left(X_{\mathsf{pix}} \mid \mathsf{wet}\right) = \\ &= \sum_{i=1}^{N_p} \frac{C(d)a_iP_i}{P_{\mathsf{pix}}} \left[\mathsf{Var}\left(X_i \mid \mathsf{wet}\right) + P_i\mathsf{E}\left[X_i \mid \mathsf{wet}\right]^2 + (P_{\mathsf{pix}} - P_i)\mathsf{E}\left[X_i \mid \mathsf{wet}\right]\right] \\ &- C(d)P_{\mathsf{pix}}\mathsf{E}\left[X_{\mathsf{pix}} \mid \mathsf{wet}\right]^2\end{aligned}$$

Finally, using Equation (A.15) and rearranging the terms I can write:

$$\begin{aligned}\mathsf{Var}\left(X_{\mathsf{pix}} \mid \mathsf{wet}\right) = \\ &= \frac{C(d)}{P_{\mathsf{pix}}} \left[\sum_{i=1}^{N_p} a_iP_i \left(\mathsf{Var}\left(X_i \mid \mathsf{wet}\right) + P_i\mathsf{E}\left[X_i \mid \mathsf{wet}\right]^2 - P_i\mathsf{E}\left[X_i \mid \mathsf{wet}\right]\right)\right] \\ &+ C(d)P_{\mathsf{pix}}\left[\mathsf{E}\left[X_{\mathsf{pix}} \mid \mathsf{wet}\right] - \mathsf{E}\left[X_{\mathsf{pix}} \mid \mathsf{wet}\right]^2\right] \qquad \text{(A.19)}\end{aligned}$$

In essence, in order to aggregate point rainfall distribution from gauges to areal distribution at on the pixel, I first aggregate the probability of rainfall occurrence P_{pix} and ratio of wet-to-wet transitions $P_{\mathsf{pix}} \cdot P_{11\mathsf{pix}}$ (Appendix A.2). I use the former to aggregate the conditional expectation of rainfall $\mathsf{E}\left[X_{\mathsf{pix}} \mid \mathsf{wet}\right]$. Both parameters are then used to aggregate the conditional variance $\mathsf{Var}\left(X_{\mathsf{pix}} \mid \mathsf{wet}\right)$. The procedure is repeated for both seasons and the four related parameters of our stochastic model (P_{11}, P_{01}, GR, GS) are calculated based on the four aggregated metrics.

For interpolation I assume that the interpolated mean and variance of conditional rainfall is a linear combination of the corresponding moments of conditional rainfall at the observation points. This allows us to apply an identical procedure as above, replacing area weights a_i with interpolation weights and setting $C(d) = 1$, as no point to areal rainfall transformation occurs.

Appendix B

Supplementary Material to Chapter 3

B.1 Derivation of the dry season streamflow CDF

The CDF of dry season streamflow Q_d is derived using the law of total probabilities

$$P_{Q_d}(q_d) = P\{Q_d \leq q_d\} = \int_{Q_0} P_{Q_d|Q_0}(q_d, q_0) p_{Q_0}(q_0) dq_0 \tag{B.1}$$

where Q_0 is the random variable representing streamflow at the beginning of the dry season.

Streamflow in any given dry season is bounded by its value at the start and at the end of the season because the receding streamflow is a decreasing function of time. The conditional CDF $P_{Q_d|Q_0}$ is therefore a piece-wise function taking the value of one for flows greater than Q_0 and of zero for flows below $Q_d(T_d, q_0)$. I first provide a detailed analysis of the bounds of the domain of $P_{Q_d|Q_0}(q_d)$ in the case of a power law recession and then describe the integration of Equation (B.1) to obtain the unconditional CDF of dry season streamflow.

B.1.1 Bounds of the conditional streamflow CDF

I examine the conditional streamflow CDF $P_{Q_d|Q_0}(q_d)$, that is the probability that the discharge governed by the deterministic recession:

$$Q_d(t) = (q_0^r - art)^{\frac{1}{r}} \tag{B.2}$$

falls below an arbitrary threshold q_d during a recession period T_d. Three regions of q_0 are immediately apparent (Figure C.1) and result in the three pieces of the conditional CDF.

(i) In the first region all streamflow values during the dry season ($t \in [0, T_d]$) lie *below* q_d for a given initial flow q_0, therefore

$$P_{Q_d|Q_0}(q_d, q_0) = 1. \tag{B.3}$$

This situation arises if

$$q_0 < q_d \tag{B.4}$$

because $Q_d(t)$ is strictly decreasing.

(ii) In the second region all streamflow values during the dry season lie *above* q_d for a given initial flow q_0, therefore

$$P_{Q_d|Q_0}(q_d, q_0) = 0. \tag{B.5}$$

Again, because $Q_d(t)$ is strictly decreasing, this situation arises for

$$q_0 \geq (q_d^r + arT_d)^{1/r} \tag{B.6}$$

However, if $r < 0$ then the values taken by streamflow at the end of the dry season must lie beneath an upper bound:

$$\max_{q_0} Q_d(T_d) = \lim_{q_0 \to \infty} Q_d(T_d) = (-arT_d)^{1/r} \tag{B.7}$$

Therefore, this second region does *not* exist (for any value of q_0) if $q_d > (-arT_d)^{1/r}$ and $r < 0$.

(iii) In the third region streamflow takes the value of q_d at some point during the dry season. This case occurs for all values of q_0 that are excluded from the two other regions. The related conditional probability can be obtained by inverting Equation (B.2):

$$P_{Q_d|Q_0}(q_d, q_0) = 1 - \frac{q_0^r - q_d^r}{arT_d} \tag{B.8}$$

The boundaries of these three regions are combined for $r > 0$ and $r < 0$ to obtain the CDF of dry season streamflow conditional on Q_0 described in Part (d).

B.1.2 Integration of the unconditional streamflow CDF

Knowing the distribution of Q_0, I apply the law of total probabilities to derive the unconditional streamflow CDF. In order to do so, I integrate Equation (B.1) within the bounds of the 3 regions described above. If $r < 0$ and $q_d > (-arT_d)^{1/r}$, the second region does not exist and Equation (B.1) integrates as:

$$\begin{aligned} P_{Q_d}(q_d) &= P\{Q_d \leq q_d\} = \int_0^\infty P_{Q_d|Q_0}(q_d, q_0) p_{Q_0}(q_0) dq_0 \\ &= \int_0^{q_d} 1 \cdot p_{Q_0}(q_0) dq_0 + \int_{q_d}^\infty \cdot \left[1 - \frac{q_0^r - q_d^r}{arT_d}\right] p_{Q_0}(q_0) dq_0 \end{aligned} \tag{B.9}$$

Inserting $p_{Q_0}(q_0) = \frac{\gamma_Q^{1+m}}{\Gamma(m+1)} exp(-\gamma_Q q_0) q_0^m$, I have:

$$\begin{aligned} P_{Q_d}(q_d) &= 1 + \frac{q_d^r}{arT_d} \frac{\Gamma_U(m+1, \gamma_Q q_d)}{\Gamma(m+1)} \\ &- \frac{\gamma_Q^{-r}}{arT_d} \frac{\Gamma_U(r+m+1, \gamma_Q q_d)}{\Gamma(m+1)} \end{aligned} \tag{B.10}$$

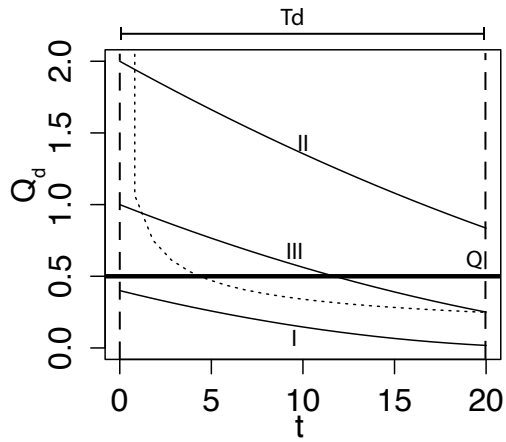


Figure B.1: Regions of $P_{Q_d \leq Q|Q_0}$. If $Q_d(0) \leq Q$ (curve I) all values of $Q(t)$ are below Q and $P_{Q_d \leq Q|Q_0} = 1$. If $Q_d(T_d) > Q$ (curve II), all values of $Q_d(t)$ are above Q and $P_{Q_d \leq Q|Q_0} = 0$. In all other cases (e.g., curve II), $Q_d(t)$ crosses Q during the dry season and $P_{Q_d \leq Q|Q_0} = 1 - (Q_0^r - Q^r)(arT_d)^{-1}$. If $r < 0$ and $Q > (-arT_d)^{1/r}$ (dotted line) $Q_d(t)$ crosses Q during the dry season for any initial streamflow value, and region 2 (i.e. $P_{Q_d \leq Q|Q_0} = 0$) does not exist.

However, if $r > 0$ or $q_d \leq (-arT_d)^{1/r}$, all three regions exist and Equation (B.1) integrates as:

$$\begin{aligned} P_{Q_d}(q_d) &= P\{Q_d \leq q_d\} = \int_0^\infty P_{Q_d|Q_0}(q_d, q_0) p_{Q_0}(q_0) dq_0 \\ &= \int_0^{q_d} 1 \cdot p_{Q_0}(q_0) dq_0 + \int_{q_{0,T_d}}^\infty 0 \cdot p_{Q_0}(q_0) dq_0 \\ &+ \int_{q_d}^{q_{0,T_d}} \cdot \left[1 - \frac{q_0^r - q_d^r}{arT_d}\right] p_{Q_0}(q_0) dq_0 \end{aligned} \tag{B.11}$$

The bound q_{0,T_d} is obtained by solving Equation (B.2) for q_0 at $t = T_d$:

$$q_{0,T_d} = (q_d^r + arT_d)^{1/r} \tag{B.12}$$

Inserting $p_{Q_0}(q_0) = \frac{\gamma_Q^{1+m}}{\Gamma(m+1)} exp(-\gamma_Q q_0) q_0^m$ I have:

$$
\begin{aligned}
P_{Q_d}(q_d) &= 1 + \frac{q_d^r}{arT_d}\frac{\Gamma_U(m+1,\gamma_Q q_d)}{\Gamma(m+1)} \\
&- \left(1+\frac{q_d^r}{arT_d}\right)\frac{\Gamma_U(m+1,\gamma_Q\left(q_d^r+arT_d\right)^{1/r})}{\Gamma(m+1)} \\
&+ \frac{\gamma_Q^{-r}}{arT_d}\frac{\Gamma_U(r+m+1,\gamma_Q\left(q_d^r+arT_d\right)^{1/r})}{\Gamma(m+1)} \\
&- \frac{\gamma_Q^{-r}}{arT_d}\frac{\Gamma_U(r+m+1,\gamma_Q q_d)}{\Gamma(m+1)}
\end{aligned}
\tag{B.13}
$$

Combining Appendix B.1.2 and Appendix B.1.2 and their respective domains, I finally obtain the expression for the CDF of dry season streamflow described in Part (d).

Appendix C

Supplementary Material to Chapter 4

C.1 Covariance of a spatially averaged process

The aim of this analysis is to explore the likely forms of a correlation structure between spatially aggregated processes, given that the underlying point-scale processes are also spatially correlated. In order to maintain tractability, the analysis will consider a strongly idealized case. While I anticipate deviations from the results in non-ideal situations, I nonetheless interpret this idealized analysis as offering insight that constrains the choice of correlation function in the TopREML analysis.

Assuming that the underlying point-scale process Y is conservative, the aggregated process y'_k related to the subcatchment S_k of gauge k can be expressed as:

$$y'_k = \frac{1}{A_k} \int_{S_k} Y(x)dx$$

where A_k is the area of S_k.

To proceed, I make the assumption that the area of the drainage areas S_k are approximately equal. While this is a strong constraint, under situations where gauges are placed near confluences and where subcatchments for a given stream ratio are adequately monitored by the gauge network, Horton Scaling ensures that the drainage areas are of a similar order of magnitude. Thus, I will take ($A_k = A \forall k$). The subcatchments are further assumed to have similar shapes and (by definition) do not overlap.

Following Cressie [64] (p. 68), the covariance between two aggregated random variables y'_k and y'_m is expressed as a function of the covariogram $C_P(\cdot)$ of the underlying point-scale process:

$$\mathsf{Cov}\,(y'_k, y'_m) = \frac{1}{A^2} \int_{S_k} \int_{S_m} C_P(\mid x_2 - x_1 \mid) dx_1 dx_2 = \int_0^\infty \nu(D) C_P(D) dD \tag{C.1}$$

where S_k and S_m are the surfaces of subcatchments k and m, and $\nu(D)$ is the probability density function of the distance between randomly chosen points within S_k and S_m – two

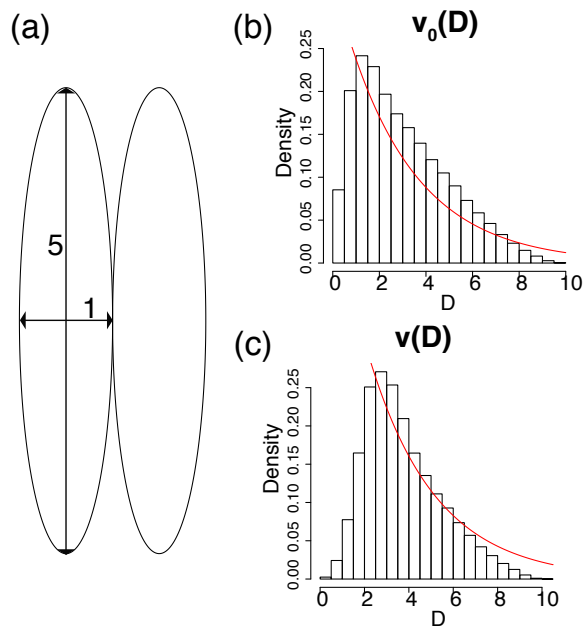


Figure C.1: The pdfs assumed in Equation (C.5) and Equation (C.2) represent well the case of adjacent ellipsoidal watersheds illustrated in subfigure (a). Subfigure (b) displays the histogram of distance between two random points *within* a watershed, overlaid by a plot of Equation (C.5) with $a_0 = 3$ and $a_D = 1/3$. Subfigure (c) displays the histogram of distance between one random point *on each* watershed, overlaid by a plot of Equation (C.2) with $a_c = 1/3$.

identical and non-overlapping shapes. Analytical expressions for $\nu(D)$ can be derived for simple geometries [e.g. 143], although complex algebraic expressions typically result. For analytical tractability I adopt a simplified expression:

$$\nu(D) = \begin{cases} a_0 exp\left(-a_D D + a_c c\right) & \text{if } c - D_1 \leq D \leq c + D_2 \\ 0 & \text{otherwise} \end{cases} \tag{C.2}$$

which approximates distance frequency function of adjacent elliptical subcatchments, as shown in Figure C.1. In Equation (C.2) the parameters a_0, $a_D > a_c$, D_1 and D_2 are positive functions of A, and c is the distance between the centroids of the subcatchments.

I also assume that the underlying point-scale process is second-order stationary and follows an exponential correlation function:

$$C_P(D) = \sigma_p^2 exp\left(-a_p D\right) \tag{C.3}$$

where σ_p^2 and a_p are respectively the point variance and spatial range of the process.

Inserting Equation (C.2) and Equation (C.3) into Equation (C.1) allows the covariance of the two spatially aggregated random variables to also be expressed as an exponential

function of the distance c between their supports

$$C_A(c) = \xi\sigma^2 exp\left(-\phi c\right)$$

where $\xi\sigma^2 = \frac{\sigma_p^2 a_0}{a_p+a_D}\left[exp(a_p D_2 + a_D D_2) - exp(-a_p D_1 - a_D D_1)\right] > 0$ and $\phi = a_p + a_D - a_c > 0$. This exponential form was adopted in the covariance derivation in the main text.

I note that within this analysis, the spatial aggregation of the point-scale process creates a nugget variance arising from spatial correlation scales smaller than the subcatchments. The nugget variance can be derived (for this idealized case) by considering the average covariance of points within the catchments:

$$\mathsf{Cov}\left(y'_k, y'_k\right) = \frac{1}{A^2}\int_{S_k}\int_{S_k} C_P(\mid x_2 - x_1 \mid)dx_1 dx_2 = \int_0^\infty \nu_0(D)C_P(D)dD \tag{C.4}$$

where $\nu_0(D)$ now represents the pdf of the distance between two randomly selected points within S_k:

$$\nu(D) = \begin{cases} a_0 exp\left(-a_D D\right) & \text{if } 0 \leq D \leq D_0 \\ 0 & \text{otherwise} \end{cases} \tag{C.5}$$

where D_0 is the maximum distance between two points within S_k. Again, inserting Equation (C.5) and Equation (C.3) into Equation (C.4), I get the nugget variance resulting from spatial aggregation:

$$C_{A,0} = \frac{\sigma_p^2 a_0}{a_p + a_D}\left[1 - exp(-a_p D_0 - a_D D_0)\right]$$

Therefore, under the aforementioned assumptions, catchment scale variance parameters σ^2 and ξ in Equation (4.6) can be expressed in terms of point scale parameters:

$$\sigma^2 = \frac{\sigma_p^2 a_0}{a_p + a_D}\left[1 - exp(-a_p D_0 - a_D D_0)\right]$$
$$\xi = \frac{exp(a_p D_2 + a_D D_2) - exp(-a_p D_1 - a_D D_1)}{1 - exp(-a_p D_0 - a_D D_0)}$$

C.2 Propagation of runoff frequency in a stream network

I describe runoff occurrence as a binary random variable taking the value of 1 if an increase in daily streamflow occurs and 0 otherwise. If runoff events are uncorrelated in time, the random variable follows a Bernouilli distribution with frequency λ. At a given gauge on a given day, the random variable takes a value of 0 if *all* of the upstream gauges take a value of 0.

In a simple situation with two upstream sub-basins described by the random variables X and Y, the frequency P_Z of the random variable $Z = max(X, Y)$ can be described as:

$$1 - P_Z = P_{!X,!Y} = P_{!X} P_{!Y|!X} = P_{!X}(1 - P_{Y|!X}) = (1 - P_X)(1 - P_{Y|!X})$$

where $!X$ stands for the event $X = 0$. Applying the law of total probabilities to substitute $P_{Y|!X}$ gives:

$$1 - P_Z = (1 - P_X)\left(1 - \frac{P_Y - P_X P_{Y|X}}{1 - P_X}\right)$$

The covariance of X and Y can be derived as:

$$\mathsf{Cov}\,(X, Y) = \mathsf{E}\,[XY] - \mathsf{E}\,[X]\,\mathsf{E}\,[Y] = P_X P_{Y|X} - P_X P_Y$$

with $\mathsf{E}\,[XY] = 0 \cdot P_{!X,!Y} + 0 \cdot P_{!X,Y} + 0 \cdot P_{X,!Y} + 1 \cdot P_{X,Y} = P_X P_{Y|X}$. Finally, substituting $P_X P_{Y|X}$ for the covariance expression, yields:

$$1 - P_Z = (1 - P_X)\left(1 - \frac{P_Y - [\mathsf{Cov}\,(X, Y) + P_X P_Y]}{1 - P_X}\right) = (1 - P_X)(1 - P_Y) + \mathsf{Cov}\,(X, Y)$$

Extending the above derivation to multiple sub-basins and neglecting the covariance term leads to a linear relation between runoff frequencies at gauge i and at upstream gauges in the following form:

$$ln(1 - \lambda_i) \approx \sum_{k=i}^{k \in \mathsf{UP}_i} ln(1 - \lambda_k)$$

Thus, if runoff pulses occur independently for each sub-basin, TopREML can be applied to $ln(1 - \lambda)$ (setting $a_k = 1$), to estimate runoff frequency at ungauged sites.

C.3 Computational Performance of TopREML

An algorithmic chart of TopREML, as implemented in the provided script, is presented in Figure C.2. IDAs and the topology of the stream network are extracted from the nested catchment using differential overlay. TopREML uses the BFGS algorithm [242] to maximize the restricted log likelihood, with the option of using a stochastic optimization algorithm (Simulated Annealing,[25]) if a non-differentiable (e.g., spherical) covariance function is selected.

A resampling analysis was performed on Austrian dataset to evaluate runtime as a function of the input data complexity, correlation structure and choice of interpolation method (TopREML versus Top-Kriging). We randomly selected one validation gauge, and resampled the remaining gauges randomly (no repetition) to generate a given prediction set size. The resampled gauges were used to estimate summer flow at the validation gauge using

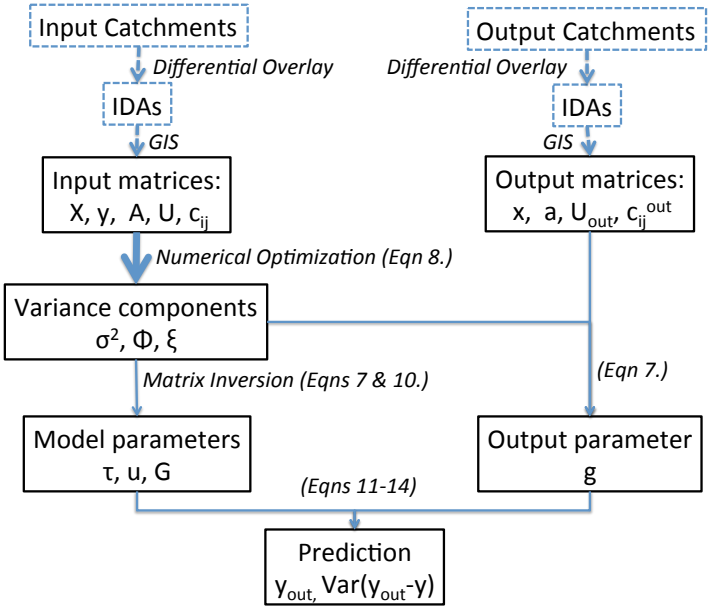


Figure C.2: Algorithmic chart of the provided TopREML implementation. Dashed frames and arrows represent vector data and operations and the bold arrow represents the step requiring numerical optimization. The complexity of the computational tasks represented by the remaining plain arrows is driven by matrix inversion, which is of polynomial complexity. In the figure, X is a matrix of observed covariate and y a vector of outcomes measured at the available gauges, as defined in Eqn. **??**; x is a vector of identical covariates observed at the prediction location. A, U and c_{ij} are matrices of relative catchment areas, network topology and inter-centroidal distances of the available gauges, as defined in Eqn 4.6; a, U_{out}, and c_{ij}^{out} are equivalent matrices for the prediction location. σ^2, ϕ and ξ are estimated variance parameters as defined in Eqn 4.3; τ, u and G are the estimated fixed and random effects (Eqn 4.10) and variance-covariance matrix (Eqn 4.7); g is the estimated covariance at the prediction location (used in Eqn 4.11) . Finally, y_{out} and $\mathsf{Var}\,(y_{out} - y)$ are the predicted outcome and the related prediction variance.

TopREML and Top-Kriging, and assuming firstly an exponential and secondly a spherical variogram. In each case, relative error and runtime were recorded. This process was repeated 200 times for each size of prediction set. Results (shown in Figure C.3) indicate that the gradient-based optimization algorithm used by TopREML for the differentiable (i.e. exponential) variogram reduces the computation runtime by an order of magnitude, relative to the implementation of Top-Kriging in the rtop package. This computational advantage vanishes if a non-differentiable (i.e. spherical) variogram must be used, which requires stochastic optimization. The results also indicate that the relative computational performance of TopREML

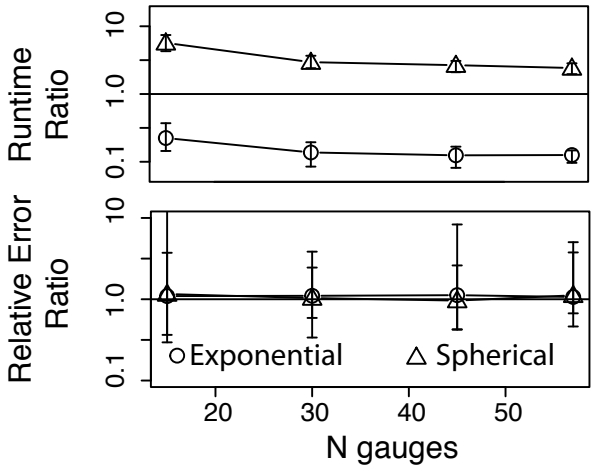


Figure C.3: Runtime resampling analysis. Leave-one-out cross-validation results for Austrian summer flow when resampling a subset of the training gauges. Computational performances are represented as the ratio of runtimes for TopREML against Top-Kriging. Prediction performances are represented as the ratio of relative errors. TopREML performances when using gradient based and stochastic optimization algorithms are represented as circles and triangles respectively. Points represent the median value and error bars represent 90% confidence intervals over 200 repetitions.

improves with the number of gauges, while its predictive performance remains constant and approximately equivalent to Top-Kriging.

Appendix D

Supplementary Material to Chapter 5

D.1 Getting price and demand from NLSS data

The NLSS dataset does not include direct data on electricity consumption and unit price, but provide household level estimates of monthly electricity expenditure (Exp), the number of common electric appliances owned[1] and the number of rooms in the dwelling (NRooms), which is related to the energy required for electric lighting. The relation between electricity expenditure and appliance ownership can be modeled as

$$\mathsf{Exp}_{iv} = \pi_v \sum_{a=0}^{A} \gamma_a n_{ai} + \varepsilon_{iv},$$

where π_v is the unit price per unit of power capacity enforced in village v, γ_a the average wattage of appliance type a and n_{ai} the number of these appliances owned by household i. $\varepsilon_{iv} \sim \mathcal{N}(0, \sigma_i^2)$ is a normally distributed error at the household level. I assume that γ_a is constant across households and villages, and that π_v is iid across villages (but constant within the villages). I wish to estimate γ_a. These assumptions allow rewriting the expression as

$$\mathsf{Exp}_{iv} = (u_v + \tilde{\pi}) \sum_{a=0}^{A} \gamma_a n_{ai} + \varepsilon_{iv} = \sum_{a=0}^{A} \alpha_a n_{ai} + u'_v + \varepsilon_{iv}$$

where $\alpha_a = \tilde{\pi}\gamma_a$ and where $u'_v = u_v \sum_{a=0}^{A} \tilde{\pi}\gamma_a n_{ai} \sim \mathcal{N}(0, \sigma_v^2)$ is a village level error that I assume orthogonal to ε_{iv}. After estimating the model using village fixed effects and removing irrelevant appliances, household demand ($\hat{D}_i$) and village unit price ($\hat{P}_v$) indices are obtained

[1] Radio, Camera, Fridge, Fan, Heater, TV.1, Phone.1, Computer. Although these represent the vast majority of appliance commonly found in rural Nepal, the omission of unusual specific appliances with high wattage (e.g. rice cookers) may lead to the underestimation of household power consumption.

by combining the estimated appliance wattages:

$$\hat{D}_i = \sum_{a \in A} n_{ai} \hat{\gamma}_a \qquad \hat{P}_v = \frac{\sum_{i \in v} \mathsf{Exp}_i}{\sum_{i \in v} \hat{D}_i} \tag{D.1}$$

The demand index is proportional to the actual household demand by a factor $\tilde{\pi}$ representing the (unknown) average electricity unit price, and the price index is the demeaned unit price. As a reality check I applied Equation (D.1) assuming a capacity of 100W for a TV to estimate the wattage of the appliances. Most fell within typical reported ranges (Figure D.1).

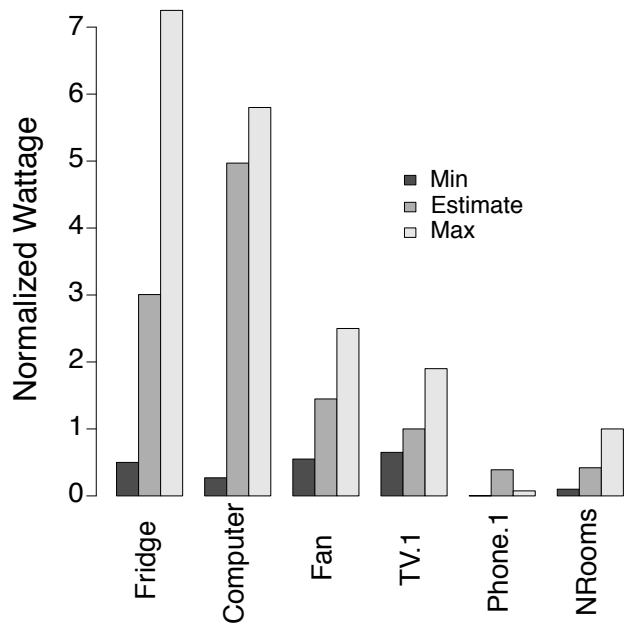


Figure D.1: Estimated appliance wattage index ('Est') compared to typical reported ranges ('Min' and 'Max') [www.lbl.gov and www.energy.gov]. Indexes are normalized by the estimated wattage corresponding to TV.

D.2 Farm-household model for electricity demand

I consider the following farm household utility maximization problem, where households seek to optimize the allocation of resources (here electricity) between production and consumption under budget constraints

$$\max_{kW_{HH}, g} U \tag{D.2}$$

$$\text{st } kW_{HH} P_{HH} + g P_g = \Pi(kW_{HH}) + w \tag{D.3}$$

where $\Pi(kW_{HH}) = kW_{HH}^{\mu} k^{\mu_k}$ is the farm production function (with production capital k and $\mu + \mu_k = 1$). $U = gkW_{HH}^{\phi}$ is the household utility function (with $\phi < 1$ related to the

quality of electricity). $P_{HH} = P_e + P_a + P_{ap}$ is the total unit cost of electricity faced by the farm-household, including the purchasing costs of the required appliances for leisure (P_a) and production (P_{ap}) purposes. P_g and g represent the unit price and consumption of a numeraire good and w are wages from market labor. kW_{HH} represents the peak electrical capacity that a household can consume at any time for leisure or production purposes. Note that in the context of unmetered electrical connections, kW_{HH} does not represent the consumed energy (in kWh), but rather the chosen electrical capacity, i.e. the 'size' of the inlet (in kW). It follows that, unlike a unit of electric energy that is consumed once used, a unit of capacity can be both used for production (during the day) and leisure (at night). Electricity demand of unmetered farm-households is therefore based on a *single* decision variable – the total demanded capacity –, whereas metered household would have to make independent choices regarding electric energy used for leisure and production. This property turns out to simplify the problem considerably, making the demand function analytically solvable in a log-log form. Similarly, P_{HH} does not represent the price paid per consumed unit of electricity, but the fixed connection fee for the chosen peak capacity.

Using Lagrange multiplyiers, I get the first order conditions:

$$\Lambda_{kW_{HH}} = \phi \frac{U}{kW_{HH}} + \lambda P_{HH} - \lambda \frac{\mu \Pi}{kW_{HH}} = 0 \tag{D.4}$$

$$\Lambda_g = \frac{U}{g} + \lambda P_g = 0 \tag{D.5}$$

$$\Lambda_\lambda = kW_{HH} P_{HH} + g P_g - w - \Pi = 0 \tag{D.6}$$

Solving the system I get

$$\phi w + (\phi + \mu) k^{\mu_k} kW_{HH}^{\mu} - (\phi + 1) kW_{HH} P_{HH} = 0 \tag{D.7}$$

Equation (D.7) does not admit an analytical solution for kW_{HH} because that variable figures in both the utility and production functions of the farm-household, meaning that the problem is not separable [19]. To solve for kW_{HH}, I assume no market for wage labor (i.e. $w = 0$). Data from the Nepal Living Standard Survey (NLSS) [166] suggest that this assumption holds for the majority (70%) of the 237 sampled households that were supplied by micro-hydropower electricity, likely because of their remoteness. In line with [e.g., 83, 204, 24], I finally get a log-log expression for the electricity demand function of the farm-household:

$$\ln kW_{HH} = \gamma + \gamma_P \ln P_{HH} \tag{D.8}$$

where $\gamma = \frac{1}{\mu - 1} \ln \frac{\phi + 1}{(\phi + \mu) k^{\mu_k}}$, and where $\gamma_P = \frac{1}{\mu - 1}$ is the price elasticity of electricity demand, which depands on μ, the scale factor of the household's production function with respect to peak electricity capacity. In line with Dias-Bandaranaike and Munasinghe [68], the model also suggests that the quality of the electricity (ϕ) and household endowment (k) affect electricity demand by shifting it for any given price.

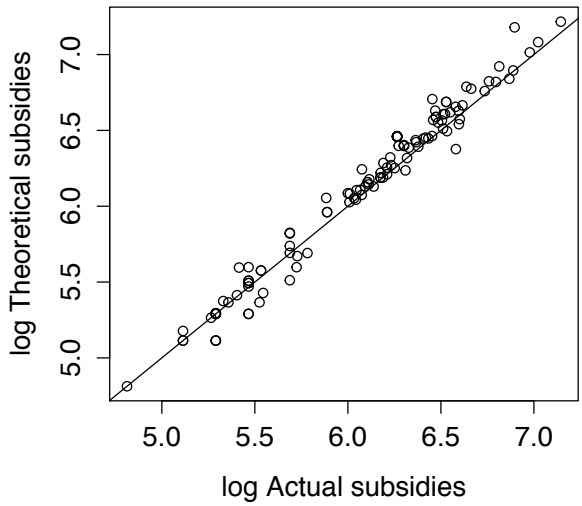


Figure D.2: Subsidy policy vs. reported subsidies by micro-hydropower scheme in the REO dataset

D.3 Subsidy policies for micro-hydropower in Nepal

Micro-hydropower subsidies in Nepal are attributed by the Alternative Energy Promotion Center (AEPC) on criteria based on the number of supplied households, infrastructure size and remoteness. The policy was updated at the end of the 2009 building season (July), and the new policy applies to all infrastructure built in 2010 and 2011 (73 schemes). The subsidy policy is well enforced in practice, as shown in Figure D.2, where the subsidies allocated by the policy are plotted against the subsidies reported in the REO dataset.

Prior to 2010[67]

- For capacities lower 5 kW: 8 kNRp per supplied household, capped at 65 kNRp per kW.
- For capacities above or equal to 5 kW: 10 kNRp per supplied household, capped at 85 kNRp per kW
- Additional subsidies for transportation costs: 1.2 kNRp per household if the site is beyond 25km from the nearest road, 3 kNRp per household if the scheme is beyond 50 km from the nearest road.

Starting in 2010[8]

- For capacities lower 5 kW: 12 kNRp per supplied household, capped at 98.5 kNRp per kW.
- For capacities above or equal to 5 kW: 15 kNRp per supplied household, capped at 125 kNRp per kW
- Additional subsidies for transportation costs: 500 NRp per kW per km if the scheme is located beyond 10 km from the nearest road, capped at 30 kNRp per kW.

D.4 Three stage least squares - GMM estimator

Following the notation in Schmidt [197], I rewrite my system of 2 equations with N observations as

$$y = Z_i\beta_i + \varepsilon_i \tag{D.9}$$

where Z_i is the matrix of right hand side variables (both exogenous and endogenous) of equation i and ε_i are random disturbances. Let us define Z and H, the matrices of the "stacked" system and instruments:

$$Z = \begin{bmatrix} Z_1 & 0 \\ 0 & Z_2 \end{bmatrix}, H = \begin{bmatrix} H_1 & 0 \\ 0 & H_2 \end{bmatrix}$$

where H_1 is the matrix of the first stage specification related to the right hand side endogenous variable of equation 1 (i.e. the exogenous variables of Z_2). Following Schmidt [197], I assume $\mathsf{rank}(Z_i^T H_i) = \mathsf{rank}(Z_i)\forall i$, $\mathsf{E}[\varepsilon] = 0$ and $\mathsf{cov}(\varepsilon) = \Sigma \otimes I_N$, where $\varepsilon = [\varepsilon_1, \ldots, \varepsilon_G]$ and where Σ is a G X G covariance matrix. I also assume that the exclusion restriction on the instruments holds for each equation: $\mathsf{plim} H_i^T \varepsilon_i = 0$. Under these assumptions, the GMM-3SLS estimator for $\beta = [\beta_1, \ldots, \beta_G]$ is unbiased and consistent and can be expressed as [197]:

$$\hat{\beta}_{GMM-3SLS} = \left[Z^T H (H^T \hat{\Omega} H)^{-1} H^T Z\right]^{-1} Z^T H (H^T \hat{\Omega} H)^{-1} H^T y \tag{D.10}$$

with $\hat{\Omega} = I_N \otimes \hat{\Sigma}$. Similarly, the variance covariance matrix of $\hat{\beta}_{GMM-3SLS}$, from which to compute standard errors on the estimates, can be expressed as:

$$\hat{C}_{GMM-3SLS} = \left(Z^T H \left(H^T \hat{\Omega} H\right)^{-1} H^T Z\right)^{-1} \tag{D.11}$$

The variance covariance matrix of the disturbances Σ (needed to compute $\hat{\Omega}$) can be estimated from the residuals e_i of a 2SLS procedure (recall 2SLS provides a consistent estimator for β)

$$\hat{\Sigma}_i j = \begin{cases} \frac{\sum_{k=1}^{N} e_{ki}^2}{N-k_i} & \text{if } i = j \\ \frac{\sum_{k=1}^{N} e_{ki} e_{kj}}{\sqrt{(N-k_i)}\sqrt{(N-k_j)}} & \text{otherwise} \end{cases}$$

Appendix E

Supplementary Material to Chapter 6

E.1 Linear Headloss Coefficient

The assumed linear head loss coefficient is obtained by considering a penstock with the following characteristics:

Diameter:	$d = 25cm$	Median value in [180]
Roughness:	$\varepsilon = 25mm$	Lightly rosted mild steel [149]
Flow Velocity:	$v = 2m/s$	Median value in [180]

Using Moody's chart [149] with a relative roughness $\varepsilon/d = 0.001$ and a Reynolds number $Re = \frac{vL}{\nu} = 500,000$, the resulting friction factor $f = 0.02$ can be inserted in the Darcy-Weisbach equation to compute a linear headloss coefficient:

$$k = f\frac{v^2}{2gd} = 0.016$$

E.2 Bivariate Ripley's K on multiple supports

Ripley's K can be used for statistical inferences on patterns of completely mapped spatial point processes, whereby the locations of all events related to these processes can be included in a predefined study area, here referred to as support. In the case of a bivariate spatial process, where events are marked with a binary attribute (here *predicted* and *observed* micro-hydro locations), the cross-function $K_{ij}(t)$ is proportional to the expected number N_j of predictions falling within a distance t of a randomly chosen observation:

$$K(t)_{ij} = \frac{1}{\lambda_j}\mathsf{E}\left[\mathsf{E}\left[N_j \in B(i,t) \mid i\right]\right] \tag{E.1}$$

where type i and type j events here indicate observed and predicted micro hydro sites. $B(i,t)$ indicates a ball of center i and radius t and λ_j is the intensity of type j events, that

is the number of events per area. It can be shown [e.g., 187] that for a homogenous Poisson process, here referred to as *complete spatial randomness* (CSR), the cross K function can be expressed as

$$K_{ij}(t) = \pi t^2. \tag{E.2}$$

The observed cross-K function is expected to be smaller than πt^2 for a regular pattern and larger if events of type i are clustered to events of type j. A widely used estimator for $K_{ij}(t)$ [70] is:

$$\hat{K}(t) = \frac{1}{\hat{\lambda}_j} \sum_{i=k} \sum_{j=l} \frac{1}{w(i, d_{i_k,j_l})} \frac{\delta(d_{i_k,j_l} < t)}{N_i} \tag{E.3}$$

with $\hat{\lambda}_j = \frac{N_j}{A}$, where A is the area of the support, $d_{i,j}$ the distance between the kth event of type i and the lth event of type j, and $\delta(\cdot)$ a Dirac delta. Edge effects are corrected by multiplying by $\frac{1}{w(i,d_{i_k,j_l})}$, the proportion of the circumference of a ball centered on i_k and of radius d_{i_k,j_l} falling inside the support area. Indeed, in order to account for points falling outside the support (and therefore not observed), the estimator is weighted inversely by the probability that such a point would be observed [187]. $\hat{K}_{ij}(t)$ can be used for statistical inference by using a Monte Carlo analysis to generate a confidence interval around $K_{ij}(t) = \pi t^2$, that is the null hypothesis that the observed pattern is a CSR process. Patterns in the point distribution are statistically significant if $\hat{K}_{ij}(t)$ falls outside the confidence interval.

The cross-K function cannot be implemented as such in the considered application because the model optimizes the location of micro hydro sites *independantly* for each community: predictions in a given community cannot be associated with observations in another community. Such cross-community associations are here prevented by considering each community as an independent support. The global cross-K function is then defined across multiple supports s as the expectation of the cross-K functions on the individual supports

$$K_{ijs}(t) = \frac{1}{\lambda_j} \mathsf{E}\left[\mathsf{E}\left[\mathsf{E}\left[N_j \in B(i,t) \mid i, s\right] \mid s\right]\right] \tag{E.4}$$

The associated estimator can be obtained by averaging Equation (E.2) over the considered supports

$$\begin{aligned} \hat{K}_{ijs}(t) &= \frac{1}{N_s \hat{\lambda}_j} \sum_{s=m} \sum_{i=k} \sum_{j=l} \left(\frac{1}{w(k, d_{k,l})} \frac{\delta(d_{k,l} < t)}{N_i} + v_m(t) \right) \\ &= \frac{1}{N_s \hat{\lambda}_j} \sum_{s=m} \left(\hat{\lambda}_{jm} \hat{K}_{ijm}(t) + v_m(t) \right) \end{aligned} \tag{E.5}$$

where $\hat{\lambda}_{jm} = \frac{N_{jm}}{A_m}$ is the intensity of type j events on support m, and $\hat{K}_{ijm}(t)$ the cross-K function estimated on support m. Non-uniform support sizes are accounted for by using the correction term:

$$v_m(t) = \begin{cases} \pi \lambda_{jm} \left(t - d_{kl}^{\max_m}\right)^2 & \text{if } t - d_{kl}^{\max_m} > 0 \\ 0 & \text{otherwise} \end{cases} \tag{E.6}$$

The correction term acknowledges the fact that no pattern can occur at a scale larger than the maximum distance $d_{kl}^{\max_m}$ between observed events in the support. This estimator is approximately unbiased and admits $\hat{K}_i js(t) = \pi t^2$ for a homogenous Poisson process, as visible in Figure 6.7, where a Monte Carlo procedure was used to generate a confidence interval for a CSR process.