Groundwater Level Dynamics and Agricultural Land-Use Change: Econometric Issues and Strategies

A THESIS

SUBMITTED IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF

DOCTOR OF PHILOSOPHY

BY

SAIF ALI

(PHD17401)

Social Sciences and Humanities

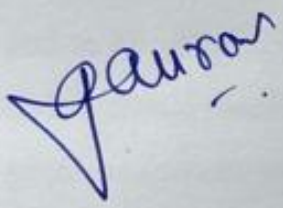
INDRAPRASTHA INSTITUTE OF INFORMATION TECHNOLOGY DELHI

NEW DELHI- 110020

17th December, 2023

THESIS CERTIFICATE

This is to certify that the thesis titled **Groundwater Level Dynamics and Agricultural Land-Use Change: Econometric Issues and Strategies**, submitted by **Saif Ali**, to the Indraprastha Institute of Information Technology, Delhi, for the award of the degree of **Doctor of Philosophy**, is a bona fide record of the research work done by him under our supervision. The contents of this thesis, in full or in parts, have not been submitted to any other Institute or University for the award of any degree or diploma.



Gaurav Arora
Thesis Supervisor
Associate Professor
Dept. of Social Science and Humanities
IIIT Delhi, 110020

Place: New Delhi

Date of Submission: 17th December, 2023

ACKNOWLEDGEMENTS

Say, "Consider this: if your water were to sink [into the earth], then who [else] could bring you flowing water?"

- The Noble Quran, 67:30

Translation: Dr. Mustafa Khattab[1]

•

I begin by expressing gratitude to Allah, who guides us from the darkness into the Light. May peace and blessings be upon the Prophet Muhammad, the noble messenger.

Thanks are due to Dr. Gaurav Arora, for his capable and compassionate mentorship and to Dr. Pankaj Jalote for his support. I owe an enormous debt of gratitude to the distinguished faculty members at ISI Delhi, the Delhi School of Economics, IIITD and other institutions who spared their extremely valuable time to teach me, advise me and review my research.

Many thanks to my family for their never-ending love and support.

(SAIF ALI)

[1]Square brackets indicate translator's insertions.

ABSTRACT

KEYWORDS: Groundwater level dynamics ; Missing data ; Spatial autocorrelation ; Agricultural land-use

A significant body of literature has pointed to a causal relationship between agricultural irrigation and groundwater depletion in India. Despite these allusions, I know of no rigorous estimation of the causal impact of cropland water demand on groundwater level change. This gap in research could be due to data availability/quality issues as well as the methodological challenge of identifying the underlying mechanisms that drive change in groundwater dynamics. In order to reconcile these challenges, I construct a unique dataset integrating satellite data products with administrative data including variables that account for climatic, hydrologic, geologic and socio-economic factors. I study three specific issues related to the identification of groundwater depletion mechanisms. First, I detect systematic or non-random missingness in administrative groundwater data due to the occurrence of "dry wells". Dry wells signify extensive depletion such that groundwater falls below the maximum depth of monitoring wells. Naive omission of dry wells can lead to severe false optimism about regional groundwater situations. I employ a set of 'observable' covariates of groundwater to predict the incidence of dry wells in an unlabelled dataset. I then utilize the prediction probabilities to quantify the statistical bias due to non-random missingness in conditional groundwater estimation models. Second, I consider the obstacles in statistical inference that arise from the fact that groundwater aquifers represent a non-exclusive common pool resource whereby the costs and benefits of resource use are shared by spatially proximate users. I employ a statistical tool known as the semivariogram to estimate spatial autocorrelation in groundwater levels. Such estimation provides empirical evidence for delineating the spatial boundaries for resource sharing within a groundwater aquifer. I then assess the impact of the spatial aquifer structure for economic policy and groundwater management science. Finally, I develop a framework for assessing the causal

impact of agricultural land-use intensification on groundwater depletion founded on a structural model that is derived from a groundwater balance equation. The identification strategy relies on a 2-stage least squares approach instrumented with spatially varying, crop-specific minimum support price (MSP) which lead to differential incentives for allocating farm acres across multiple crops and hence groundwater extraction outcomes. This work advances the study of the causal relationship between groundwater irrigation and depletion by addressing three oft-ignored econometric issues that arise in such a study. Overall, my essays bear relevance for groundwater management and policy making as well as academic research where accurate and efficient estimation of statistical moments of groundwater levels is of paramount importance.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

ABBREVIATIONS

GWL(s)	Groundwater level(s)
LULC	Land-use land-cover
GB	Gangetic Basin
UP	Uttar Pradesh
NI	North India
CGWB	Central Ground Water Board (or Central Groundwater Board)
UPGWD	Uttar Pradesh Ground Water Board (or Uttar Pradesh Groundwater Board)
MI, MIC	Minor Irrigation Census
UPGWMA	Uttar Pradesh Ground Water Management Act
SGMA	Sustainable Groundwater Management Act
USGS	United States Geological Survey
BGS	British Geological Survey
NCA	Net Cropped Area
GCA	Gross Cropped Area
CPR(s)	Common Pool Resource(s)
OLS	Ordinary least squares
GLS	Generalized least squares
MLE	Maximum likelihood estimate (or estimation)
LD	Listwise deletion
MCAR	Missing completely at random
MAR	Missing at random
MNAR	Missing not at random
IMD	Indian Meteorological Department
ISRO	Indian Space Research Organization
NRSC	National Remote Sensing Center
WRIS	Water Resource Information System
SGE	State of groundwater extraction

CHAPTER 1

SUMMARY

This work was motivated by a meeting of the author with the ex-chairman of the Central Groundwater Board (CGWB), the apex body for groundwater management in India. The meeting brought to light the issue of "groundwater depletion" in the country. Aquifers, the geological formations that store groundwater, are likened to savings accounts and the groundwater to money. In a state of equilibrium the withdrawals (discharges) of groundwater from an aquifer equal the deposits (recharges) (USGS, 2003). Groundwater depletion refers to the long-term regional decline of "groundwater levels" usually due to continual groundwater withdrawal for human use in excess of recharges (USGS, 2018a, 2003). The groundwater level (GWL) is the depth from the ground surface at which the sub-soil becomes saturated with water. It is also known as the "water table" (BGS, 2022b). Importantly for economics, it is the depth to which a pumping well must be dug or drilled in order to draw water to the ground for human use. Groundwater depletion, the gradual decline in GWLs, is associated with increasing water stress, water scarcity, drying up of wells (USGS, 2018a), increased pumping costs[1] (Narayanamoorthy, 2015), increased uncertainty of the success of drilling[2] (Nageswara Rao et al., 2009), increased rural poverty and social conflict (Sekhri, 2014; Unfried et al., 2022), reduced agricultural yields (Jain et al., 2021) and reduced scope for adaptation to climate change (Fishman, 2018). Groundwater irrigation through tube-wells gives farmers relatively greater control over irrigation schedules, protection from uncertainty in rainfall and the mitigation of the risk of exposure to drought at the cost of drilling wells, purchasing lift technology and paying for electricity (Tsur, 1990). Depletion reduces the security buffer offered by irrigation by increasing costs, risks and uncertainty.

[1]The cost of modifying a bore-well to extract water from deeper levels could be as much as 48% of the cost of the well itself (Narayanamoorthy, 2015).

[2]Farmers in Andhra Pradesh were reported to have only a 25% success rate at digging bore wells in 2007-2008 while the cost of digging was 35 rupees per foot (Nageswara Rao et al., 2009).

There is evidence that during the post-independence (1947) years in India, groundwater irrigation was a strong driver of shifts in agricultural land use towards "multiple-cropping" (Dayal, 1977; Narain & Roy, 1980; Dhawan & Datta, 1992). Multiple cropping is the practice of cultivating a piece of land more than once a year. The "net cropped area" (NCA)[3] is the total area of land sown with crops and orchards in a given year with multiple cropped lands being counted only once. The "gross cropped area" (GCA) is the total area sown with crops and orchards with multiple-cropped lands being counted as many times as there was sowing (DES, 2008). Post-Independence agricultural reforms that preceded the "Green Revolution"-era (1967-78 to 1977-1978) ushered in a thirty-year period of agricultural intensification, mostly via multiple cropping[4], fueled partially by aggressive groundwater exploitation (Mukherji, 2022; Shah, 2010; Dantwala, 1976). The agricultural reforms of the Green-Revolution era were aimed at ensuring national food security and promoted tubewell irrigation with an outlook of realizing the 'full' potential of untapped national groundwater resource. Farmers were subsequently given electricity subsidy and minimum support prices (MSP, floor prices) to promote paddy and wheat production. India's irrigated area expanded by 150% between 1950 and 1990 (Mukherji, 2022). Then in the early 1990s concerns around groundwater depletion emerged (Chandrakanth & Romm, 1990). In the Northern Plains of the Indian subcontinent, the euphoria around groundwater development in the years after independence gave way to an urgent need for groundwater resource management in the years leading up to millennium (Shah, 2010; Subramanian, 2015; Dantwala, 1976).

In the 2000s, a large and influential literature in hydrology, sustainability science and climate science links groundwater depletion to irrigation. Anuraga (2006) observes that "farmers in many areas are using groundwater faster than nature is replenishing it, causing continuous declines in the water levels" (Anuraga et al., 2006). Rodell *et. al.* (2009) conclude that "the available evidence suggests that unsustainable consumption of groundwater for irrigation and other anthropogenic uses is likely to be the cause [of groundwater depletion]" although they qualify the conclusion with the caveat that their evidence is limited (Rodell et al., 2009). Siebert observes that depletion in arid and

[3] Also known as the "Net Area Sown"

[4] In the 1960s and later years, very little new land was brought into cultivation and expansion of cultivated area was achieved via multiple cropping of existing agricultural lands (Abel, 1970).

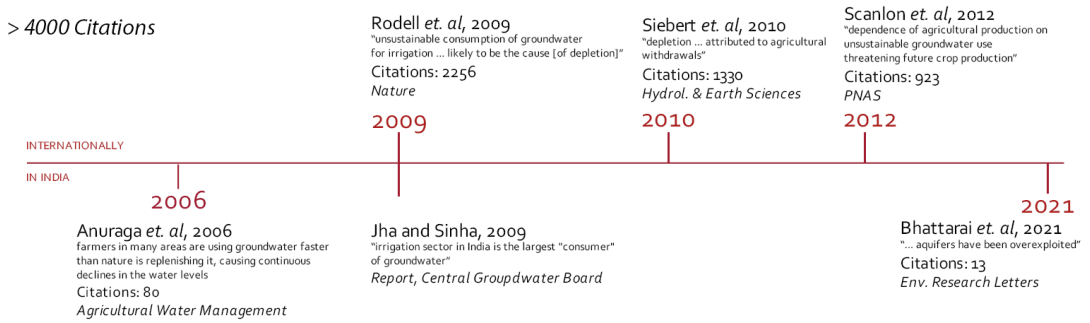


Figure 1.1: Influential papers that suggest or allude to a causal link between irrigation and groundwater depletion.

semi-arid areas worldwide can be "attributed to agricultural withdrawals" (Siebert et al., 2010). Other studies make similar remarks (Ahmed & Umar, 2009; Scanlon et al., 2012) (see Figure 1.1).

Research and policy investment in groundwater management has increased steadily since 2000 (see Figure 1.2). Over 200 papers were published on the themes of "groundwater irrigation" and "depletion" in the year 2020 and major policy initiatives for groundwater management were introduced in USA (2014; 2014a, 2014b) and India (2020; 2019) between 2015 and 2020. The "Uttar Pradesh Ground Water Management Act" (UPGWMA) was passed in 2019 and a portion of it regulates groundwater use for irrigation by farmers. There is evidence that irrigation accounts for a large percentage of groundwater use in India[5] (Siebert et al., 2010) as well as globally[6] (Wada et al., 2014) which suggests but does not confirm a causal impact. Despite several allusions to a causal link between irrigation and groundwater depletion and major research and policy investment in this area, to the best of our knowledge, no papers in economics specifically measure the causal impact of irrigation withdrawal on groundwater depletion at the regional and sub-regional scale. This is a gap in the research and policy literature.

The existence of such a gap may be partially explained by the lack of available data and partially by methodological challenges associated with econometric estimation and causal inference for GWL dynamics and irrigation water use. There is no open,

[5]Siebert (2010) has estimated that in a global inventory of over 15000 administrative units, India had the largest conjunctive use of groundwater for irrigation at 204 km^3 yr^{-1} (Siebert et al., 2010).

[6]Wada *et. al* report that agriculture is the "dominant user of water" accounting for 70% of total world consumption by volume (Wada, Wisser, & Bierkens, 2014).

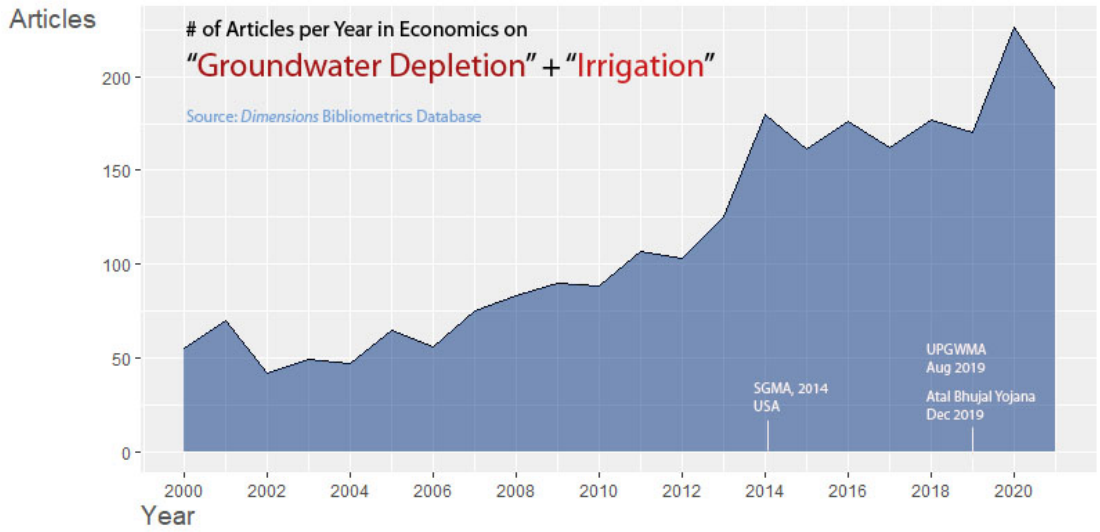


Figure 1.2: Number of research publications on the themes of "groundwater depletion" and "irrigation" and policy enactments on groundwater management, 2000-2020. Source: Dimensions bibliometric database.

public and free source that provides spatially dis-aggregated data on groundwater use for irrigation[7]. We exploit high-resolution, high-frequency, spatio-temporal, satellite sensor data on *land-use land-cover* (LULC) (NRSC, 2007) as a measure of groundwater use for irrigation. We specifically study the impact of shifts in land-use patterns from "single-cropping" to "multiple-cropping" that entail intensive irrigation water use. Our remote-sensing LULC data provides indicators of such shifts at a high spatial resolution of 56 meters[8] (ISRO, 2022). The high spatial resolution enables us to study the causal impact of agricultural land-use transitions on GWL dynamics at the level of individual groundwater monitoring wells.

The goal of this study is the econometric estimation of spatially-delineated groundwater level (GWLs) dynamics in relation to agricultural land use intensification. The study area is the Northern Indian state of Uttar Pradesh (UP) where rapid intensification of irrigated agriculture is associated with depletion of groundwater resources (Rodell et al., 2009) and the study period is 2009-10 to 2019-2020. The study area, i.e the state of UP located within the GB, is a major source of national agricultural production in-

[7]The Minor Irrigation Census (MI Census, 2013b) publicly provides irrigation water use and GWLs data aggregated to the district level.

[8]This data is obtained using the ResourceSat 1, 2 2A-AWiFS sensor and represents the highest spatial resolution for data that are available across the study period. LULC data at 10 m and 5.8 m resolutions are available only for selected years (ISRO, 2022).

cluding for staples like wheat, rice and pulses, and also cash crops like sugarcane and maize. The GB is endowed with a fertile soil cover and thick alluvial aquifers with high water bearing capacity and abundant Monsoon rainfall that increases from west to east (Muñoz Sabater, 2019). A multi-layered aquifer system underlies UP forming one of the richest groundwater repositories of the world where groundwater contributes 71% of the total irrigation water needs of the state (CGWB, 2021). The state of UP is an agrarian economy that was home to 16.5% of the total population of India in 2011 (Census of India, 2011) of which 47% directly depended on agriculture to make their living (Gulati, Terway, & Hussain, 2021).

My dissertation research addresses three important challenges for econometric estimation and inference that arise due to technological groundwater monitoring constraints (Ali & Arora, 2021), the common-pool nature of groundwater resources (Ali & Arora, 2022) and the endogeneity of agricultural land-use to local GWL.

In the second chapter, I address the question of "missing data" that arises due to monitoring constraints associated with groundwater monitoring wells (Ali & Arora, 2021). GWL is typically measured as the depth of water table from ground surface using monitoring wells owned by government agencies (CGWB, 2017; Cunningham & Schalk, 2014)[9]. A monitoring well is termed as "dry well" whenever the water table falls below its maximum floor depth (USGS, 2019). Clearly, dry wells cannot measure actual GWL at a given location and so well-dryness manifests as missingness in administrative GWL data publications (Hora et al., 2019)[10]. I formally characterize information loss due to dry groundwater monitoring wells (UPGWD, 2020) and evaluate the consequence(s) of "missing values" in GWL data on statistical inference in the context of economic policy and groundwater management science. I argue that *missingness due to dry wells* is non-random, and therefore non-ignorable[11] with regards to attain-

[9]Remote sensing technology has been recently employed for GWL monitoring, however these data have relatively lower spatial resolution (Ali et al., 2021) and suffer from substantial measurement errors (Long et al., 2016; Bhanja, Mukherjee, Saha, Velicogna, & Famiglietti, 2016). Most scientific and policy literature relies on wells based GWL monitoring.

[10]In Uttar Pradesh, India (for example) I found over one-third GWL observations (i.e., >50,000 observations) were missing or not reported during 2009-2019 period (UPGWD, 2020), and that dry wells constituted a major cause of missing GWL data (see Figure 2.1).

[11]Technically, if we could find a vector of explanatory variables that explains the occurrence of missingness perfectly, it would then be classified as "random" missingness. In practice, finding such and explanation is not possible. It is also not possible to verify such an empirical explanation against the truth.

ing consistent inference on the groundwater situation in an area. When a well dries up the corresponding GWL record goes missing. As more wells dry up, as might happen with increasing water scarcity (Perrone & Jasechko, 2017a; Varghese et al., 2013; USGS, 2018b), and more GWL observations are removed from a given sample, the inference about worsening groundwater depletion using the surviving sample (i.e., excludes missing values) will be systematically biased. A greater truncation of worsening GWL instances in a locality could lead to *false optimism* about resource availability for its utilization.

I propose a novel, integrated framework to assess the impact of truncation of missing data on estimation and inference of the means and conditional means of GWLs from real-world datasets. The centerpiece of the framework is a logistic regression that explains and predicts the incidence of missing values due to dry wells by a vector of carefully chosen, spatially delineated covariates. The model facilitates the characterization of systematic (non-random) variation in missingness, prediction of missingness due to dry wells, the estimation of the contribution of each factor affecting missingness and the estimation of the biases that result from truncation of missing values. It also provides confidence intervals for the estimated biases. It accounts for the typical factors influencing dryness of wells like rainfall, temperature, cropping and urbanization but we also include variables that capture groundwater "runoff" and distance from surface water bodies which are known to influence GWLs but are typically not accounted for; further all of this is conducted at the dis-aggregated level of individual monitoring wells and separately for each well type (DEEP and SHALLOW).

The issue of information loss due to dry monitoring wells bears relevance to groundwater management where GWL data are employed for the estimation of regional groundwater stock, extraction and recharge levels (BGS, 2022a; CGWB, 2021; USGS, 2013b; Qiu et al., 2010; C Koreimann & Vogel, 1996). It is also relevant to applied economics research where (for example) GWL data are used to measure the impact of groundwater on agricultural growth (Jain et al., 2021), poverty (Sekhri, 2014), technology adoption (Li & Zhao, 2018) and social welfare issues (Sayre & Taraz, 2018). Information loss due to dry wells raises an important concern for groundwater monitoring in India and elsewhere under future climate scenarios that project higher incidence of droughts

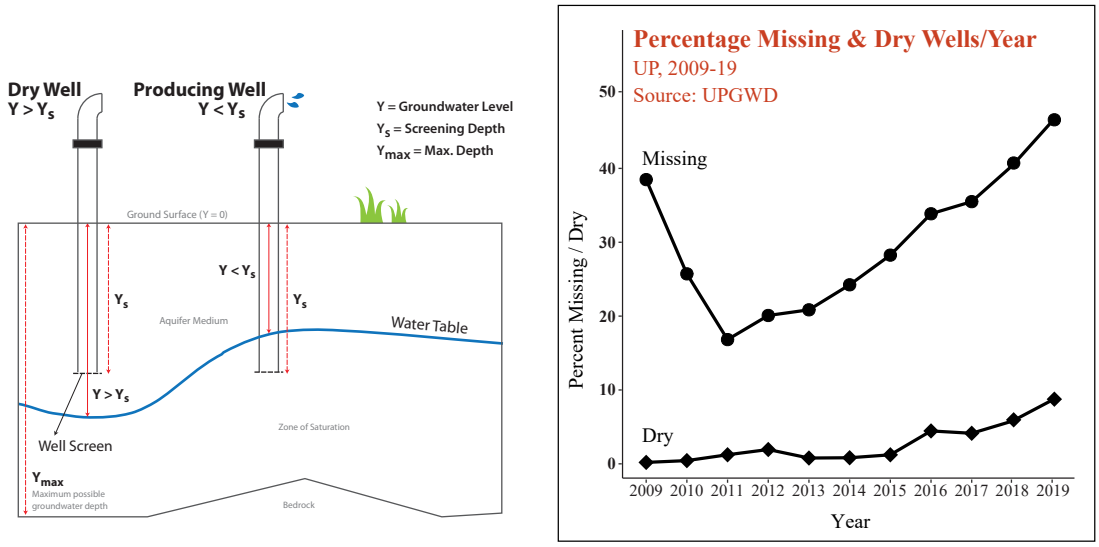


Figure 1.3: Data truncation due to a dry well (left) and the percentage of missing data and dry wells by year for Uttar Pradesh (UP), North India, 2009-19 (right).

(Arora & Ali, 2022).

Groundwater aquifers have the nature of a non-exclusive, common pool resource (CPR) (Gardner et al., 1990; Ostrom, 1990) whereby the costs and benefits of resource usage are shared by spatially proximate users (Pfeiffer & Lin, 2012; Edwards, 2016). Groundwater pumping creates a "spatial externality" so that the extraction by one user influences the GWLs experienced by nearby users (Pfeiffer & Lin, 2012). In the third chapter (2022), I describe a geostatistical spatial model (N. A. Cressie, 1993) for GWLs and apply it to characterize the extent and structure of "spatial auto-correlation" in GWLs. Spatial auto-correlation refers to the correlation between the observed value of GWL at one location to that of *its own* (auto) values at nearby locations (Anselin, 2013). Figure 1.4 shows GWLs in the Indo-Gangetic basin aggregated to the district level (left panel) and at individual monitoring wells in the sub-region of Western UP (right panel). Note that GWLs in the basin exhibit spatial auto-correlation at both scales. I use a functional instrument called the "semivariogram", to model the structure of spatial auto-correlation in GWLs towards two important tasks. First, to estimate the range (or distance) within which groundwater users might influence each other. Second, to spatially delineate hot-spots of resource depletion where welfare gains from management are expected to be the greatest (Edwards, 2016). I use simulation to demonstrate how spatial auto-correlation (or spatial dependence) between GWL observations violates the classical assumptions of the Gauss-Markov theorem needed for unbiased estimation of

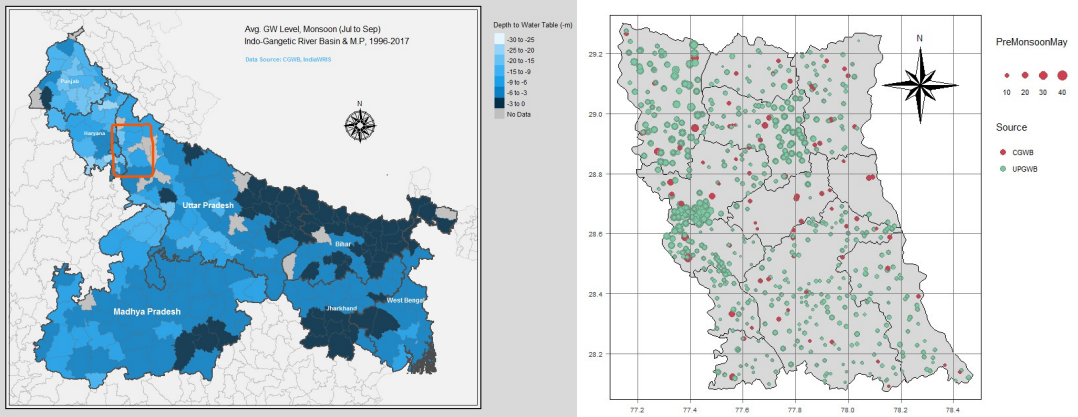


Figure 1.4: Spatial Auto-Correlation at the district (left) and well (right) scale.

linear regression coefficients via ordinary least-squares (OLS).

Identification of land use impacts on GWL will rely on a stark shift in agricultural land use changes in the part of the Gangetic Basin (GB) covering UP due to an increase in multiple cropping acreage during the study period (Arora et al., 2019). Such land use changes are driven, at least partially by, access to groundwater irrigation supported by conducive policy interventions like highly subsidized electricity and distribution of public tube wells (Pant, 1994). Significant increase in minimum support price (i.e., floor price) for crops procured by the government and large-scale agricultural loan waivers circa 2009 are also potential drivers of land use intensification in the GB. In the fourth chapter, I exploit exogenous variation in land use changes through an instrumental variables regression framework to ultimately establish the causal impact of land use changes on GWLs.

CHAPTER 2

DRY WELLS: A MISSING DATA CONUNDRUM FOR GROUNDWATER MANAGEMENT

2.1 Introduction

Groundwater resource management has assumed global importance amidst significant resource depletion in important food producing regions like the Indian sub-continent, China and USA (CGWB, 2021; CGWB, 2019; Schoengold & Brozovic, 2018; Sears et al., 2018; Famiglietti, 2014; Konikow & Kendy, 2005; Rodell et al., 2009). Effective resource management requires adequate and continuous monitoring of regional groundwater stock, extraction and recharge levels (BGS, 2022a; CGWB, 2021; USGS, 2013b; Qiu et al., 2010; C Koreimann & Vogel, 1996). Major policy initiatives in India (2019; 2020) and the United States (2014; 2014a, 2014b) emphasize the utility of groundwater level (GWL) monitoring data for measuring spatio-temporal changes in groundwater resources (Asoka et al., 2017; Bhanja et al., 2016) and for implementing policy efforts that are aimed at efficient resource management (Zirulia et al., 2021; Ministry of Jal Shakti, 2020; UPGWD, 2019; CGWB, 2019b; Schoengold & Brozovic, 2018; Dickinson, 2014; Pavley, 2014a, 2014b). GWL datasets are employed in a wide-range of models, e.g., for groundwater balance (i.e., difference between resource availability and extraction) estimation in the hydrology literature, and to model agricultural growth as function of groundwater access in applied economics papers[1].

GWL is typically measured as the depth of water table from ground surface using monitoring wells owned by government agencies like the Central Groundwater Management Board in India and Water Resources Departments in the United States (CGWB,

[1]Hydrologists use GWL data to evaluate regional balance between groundwater recharge and extraction (Jukić & Denić-Jukić, 2009), assess hydrological response to climate change (Scibek & Allen, 2006) and model water flows in urban areas (Fletcher, Andrieu, & Hamel, 2013). Economists use GWL data to measure the impact of groundwater on agricultural growth (Jain et al., 2021), poverty (Sekhri, 2014), technology adoption (Li & Zhao, 2018) and social welfare issues (Sayre & Taraz, 2018).

2017; Cunningham & Schalk, 2014)[2]. Monitoring wells have a finite depth from the ground surface, known as the maximum floor depth. A monitoring well is termed as "dry well" whenever the water table falls below its maximum floor depth (USGS, 2019). Figure 2.1 schematically illustrates a dry well. Clearly, dry wells cannot measure actual GWL at a given location and so well-dryness manifests as missingness in administrative GWL data publications (Hora et al., 2019). In Uttar Pradesh, India (for example) we found over one-third GWL observations (i.e., >50,000 observations) were missing or not reported during 2009-2019 period (UPGWD, 2020), and that dry wells constituted a major cause of missing GWL data. This paper aims to formally characterize information loss due to dry wells (UPGWD, 2020; Perrone & Jasechko, 2017a) and evaluate the consequence(s) of "missing values" in GWL data on statistical inference in the context of economic policy.

Table 2.1 provides evidence that the existing studies frequently ignore or overlook missing GWL values both in economics (Bhattarai et al., 2021b; Smith, 2018; Hrozencik et al., 2017; Sekhri, 2014; Pfeiffer & Lin, 2014, 2012) and non-economics literature (Asoka et al., 2017; Bhanja et al., 2016; Meghwal et al., 2019; Sarkar et al., 2020; MacDonald et al., 2016; Panda & Wahr, 2016; Asoka & Mishra, 2019; J. Das et al., 2020). Moreover, missingness in administrative GWL data remains out-of-bounds in India's 'Groundwater Estimation Committee (GEC)' reports, see for example (CGWB, 2009). In the United States, the 'National Framework for Groundwater Monitoring' stipulates that missing GWL values are not artificially imputed due to heterogeneity in groundwater aquifer systems (USGS, 2013b, p.135). It is interesting to note that while hydrologists have developed imputation methods for missing groundwater data since Dax (1985)[3], the groundwater management literature in economics and public policy domains have paid little or no attention to this important matter.

[2]Remote sensing technology has been recently employed for GWL monitoring, however these data have relatively lower spatial resolution (Ali et al., 2021) and suffer from substantial measurement errors (Long et al., 2016; Bhanja et al., 2016). Most scientific and policy literature relies on wells based GWL monitoring.

[3]These include classical imputation (Semiromi & Koch, 2019; Kenda, Koprivec, & Mladenić, 2018; Dax & Zilberbrand, 2017), geostatistical prediction (Chung, Senapathi, .S, & Prasanna, 2019), spatial interpolation, (Parasyris, Spanoudaki, Varouchakis, & Kampanis, 2021) and more recently machine learning based approaches (Evans, Williams, Jones, Ames, & Nelson, 2020; A. Mukherjee & Ramachandran, 2018). Please see Table 2.2 (Xiong, Abhishek, Guo, & Kinouchi, 2022; Huang, Cao, Yu, Liu, & Wang, 2021).

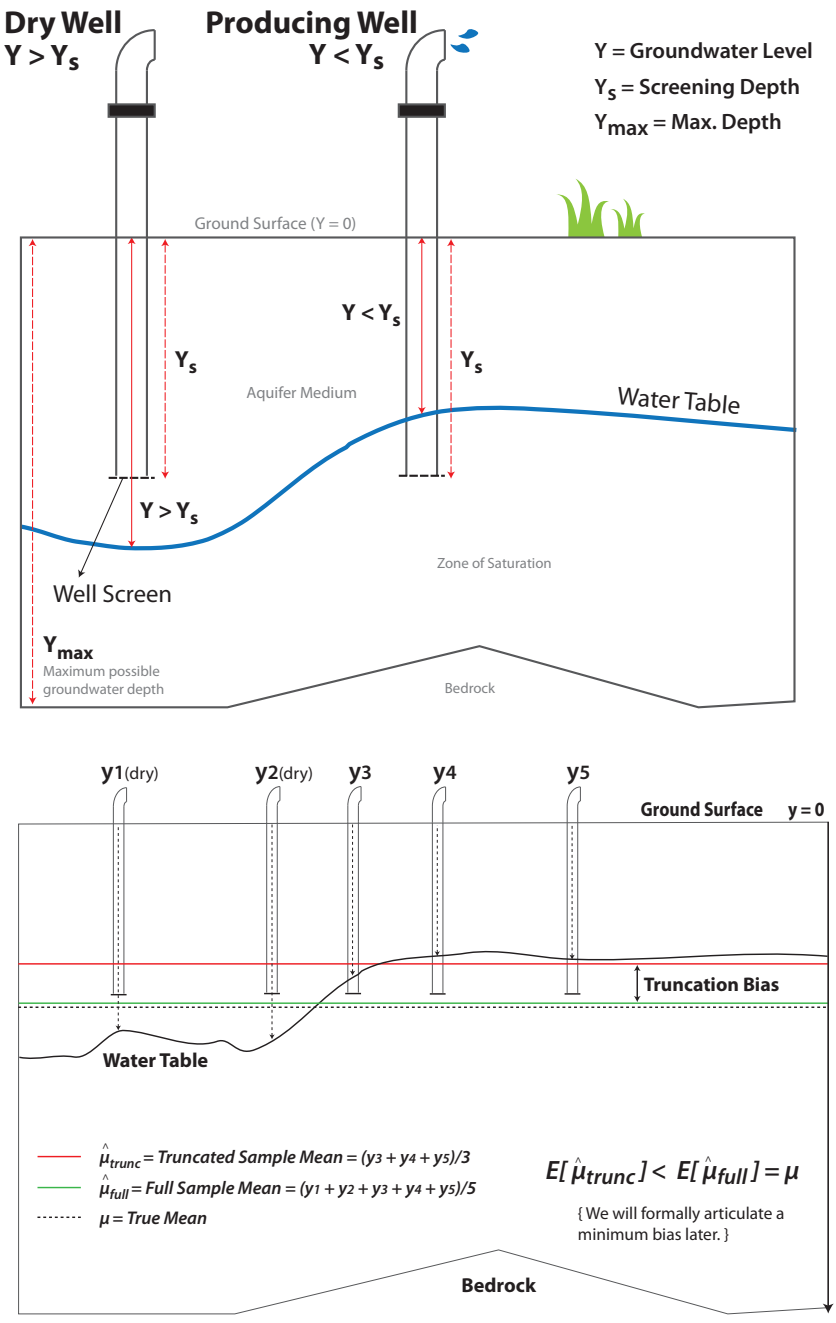


Figure 2.1: **Dry wells and spatial mean estimation.***Top panel:* A schematic diagram showing a dry and producing (non-dry) well. *Bottom panel:* Example of spatial mean estimation while ignoring dry wells.

Table 2.1: Approaches adopted in the literature for handling missing GWLs data .

Non-economics disciplines

Study (Author-Date)	Model/Statistic	Spatial Aggregation	Missing Data Approach
Xiong et al. (2022)	-	Well	ML-based imputation
Huang et al. (2021)	Morlet's wavelet transform	Well	Spline interpolation
Asoka et al. (2017)	Mann Kendall Trend test, Sen's Slope Estimator, linear regression	Well	List-wise Deletion (LD)
Bhanja et al. (2016)	Linear regression, non-linear trend analysis	Basin	LD
Panda and Wahr (2016)	Non-parametric Mann Kendall test	Basin	LD
Zaveri et al. (2016)	Decadal sample mean of GWL	District	None mentioned

Economics

Study (Author-Date)	Model/Statistic	Spatial Aggregation	Missing Data Approach
Bhattarai et al. (2021b)	Ordinary least squares (OLS), GWL as regressor	District	LD
Smith (2018)	OLS, GWL as dependent variable	Well	None mentioned
Hrozencik et al. (2017)	OLS, GWL as regressor	Well	None mentioned
Sekhri (2014)	2-stage least squares (2SLS), GWL as regressor	Village	None mentioned
Pfeiffer and Lin (2014)	Probit/Logit, GWL as regressor	Farm/County	None mentioned
Pfeiffer and Lin (2012)	2SLS, GWL as dependent variable	Well	None mentioned

Table 2.2: Approaches proposed in the literature for handling missing GWL data.

Study (Author-Date)	Discipline	Spatial Aggregation	Missing Data Approach
Chung *et. al* (2019)	Hydrology	Well	Geostatistical modeling
Semiromi & Koch (2019)	Hydrology	Well	Imputation
Kenda (2018)	CSE and others	Well	Imputation
Dax & Zilberbrand (2017)	Hydrology	Well	Imputation
Escareño *et. al* (2022)	Hydrogeology	-	Optimal well placement
Parasyris *et. al* (2021)	Hydroinformatics	-	Genetic algorithms, sp. interpolation
Evans *et. al* (2020)	Environmental Sc.	-	ML based imputation
Mukherjee *et. al* (2018)	Hydrology	-	ML based imputation
Dax (1985)	Hydrology	Well	Interpolation

We argue that *missingness due to dry wells* is non-random, and therefore non-ignorable with regards to attaining consistent inference on the groundwater situation in an area. This is because the data generating process underlying missing GWL values due to well-dryness is itself a function of water table depth (Rubin, 1976). In other words, when a well dries up, i.e., the water table drops below its maximum floor depth, the corresponding GWL record goes missing. And as more wells will dry up, which is a marker of water scarcity (Perrone & Jasechko, 2017a; Varghese et al., 2013; USGS, 2018b), and more GWL observations are removed from a given sample, the inference about worsening groundwater depletion using the surviving sample (i.e., excludes missing values) will be systematically biased. In fact, a greater truncation of worsening GWL instances in a locality could lead to *false optimism* about resource availability for its utilization (Arora & Ali, 2022).

In order to formalize the impact(s) of non-random missingness on regional groundwater assessment we follow Rubin's missing data theory (Mealli & Rubin, 2015). Specifically, we posit missing GWL observations to be drawn from binary-valued random variable which equals 1 if GWL record is missing and 0 otherwise. Then, the sample GWL estimates like the spatial mean over a village or state will depend not only on the probability distribution underlying *observed* GWL data but also the probability distribution underlying *missing* values (Rubin, 1976). Therefore, naive deletion of missing GWL records when conducting statistical inference would amount to ignoring the probability distribution of missing data .

We address whether it is appropriate to perform sample-based inference on the surviving sample after the missing values have been deleted. Rubin had shown that deleting missing values would not affect statistical inference on some data parameter only when the missing values are "missing completely at random" (i.e., distribution parameters of the sample and those of the missingness process are "distinct" entities) (Rubin, 1976). We will evaluate whether dry wells-led missing data are non-random based on a statistical comparison between the mean values of relevant covariates (like rainfall, urbanization, etc.) in the local neighborhood of missing and non-missing GWL observations (Little & Rubin, 2019; Rubin, 1976; Mealli & Rubin, 2015). Then, we will estimate a multivariate logistic regression to establish conditional dependence beween probability

distribution of missing GWL records and groundwater stressors like drought, irrigation and urban development.

Groundwater resource management through policy making and/or research inquiries rely on the estimation of regional GWL situation. However, such efforts have paid little attention to the missing data problem. We focus on two common estimation strategies namely unconditional mean estimation via *spatial aggregation* over some administrative unit (e.g., village, block[4], district or state, or a hydrological unit like watershed/basin) (CGWB, 2018; India Observatory, 2022; MI Census, 2013b) in a given year, and conditional mean estimation via *linear regression*. While individual well-level GWL data are publicly available in India, most existing applications use spatially aggregated versions. Therefore, accurate spatial aggregation is a central concern for groundwater policy where resources are typically managed in an administrative unit (e.g., blocks in India (CGWB, 2021)). A bias due to unaccounted missingness due to dry wells is likely implicit or "built-in" when using spatially aggregated datasets. We thus consider characterization of missingness at the well-level as an important concern.

Linear regression models have been previously employed to estimate the impact of GWL values on agricultural growth (Jain et al., 2021), poverty (Sekhri, 2014), technology adoption (Li & Zhao, 2018), social welfare (Sayre & Taraz, 2018) (i.e., GWL is a regressor), as well as the impact of irrigated agriculture, rainfall and policy interventions on regional GWL situation (i.e., GWL is a dependent variable) (Kirby et al., 2015; Guzmán, Paz, Tagert, Mercer, & Pote, 2018) (Asoka et al., 2017; Bhanja et al., 2019) (Bhanja et al., 2017). It is common practice to "list-wise" delete missing GWL revords, please see Table 2.1, wherein missing values of the dependent variable are deleted *along* with all associated covariates. Therefore, list-wise deletion (LD) results in a limited-dependent variable model whenever truncated GWL data are modeled as a function of covariates (Tobin, 1958). The standard (or type-1) Tobit model for regression with truncated data is equivalent to explicitly accounting for the process that causes missing data as stipulated by Rubin (Amemiya, 1984; Rubin, 1976). We use a simulations-based approach to show that for truncated regression, estimated coefficients are biased even for exogenous regressor like rainfall and that the magnitude of the bias depends on the

[4]The assessment unit in India is called a "block". A block is sub-district administrative unit used for implementation of government programs in rural communities at the Gram Panchayat level."

coefficient of correlation between GWL and the regressor of interest.

Hora et al. (2019) speculated that that ignored missing values due dry wells to dry wells give rise to survivor bias in statistical models, however to our best knowledge no previous study has analyzed or measured a bias in statistical inference that may arise due to naive deletion of missing data. This paper contributes to natural resource management literature in the following manner. First, we establish well-dryness as a formal mechanism of non-random missingness in GWL data based on the missing data theory. Second, we present a method to predict dry well-related missingness from non-experimental data with a multivariate logistic regression model. We use the estimated probabilities of missingness due to dryness to demonstrate the bias introduced from LD of missing values on the estimate of the mean and the coefficients of regression on the condition mean of GWLs. We focus on an empirical case study of GWL data in Uttar Pradesh (UP) state in North India. We discuss the significance of our results for groundwater management and policy making in India.

The rest of the chapter is organised as follows; Section 2.2 will present a background of groundwater management problem in UP in the presence of missing data. Section 2.3 will provide an outline of our methodology, followed by a discussion of our main results (Section 2.4) and some concluding remarks (Section 2.5).

2.2 Study region, GWL monitoring and missing data

We explore groundwater management issues for the northern Indian state of Uttar Pradesh (UP). UP is India's most populous state with 20 million inhabitants (16.6% population) that spans 243,000 square kilometers[5] (7.3% landmass) in geographic area (Census of India, 2011). Here we argue that our regional focus exhibits a general representation of groundwater management problem in the presence of missing data. UP is endowed with the fertile and water-rich Gangetic basin, which supports a large agricultural sector that employs 47% of the state's population (Gulati et al., 2021) for the cultivation of staples like wheat, rice and pulses, and cash crops like sugarcane and maize. Besides sustaining

[5]Among countries the size is comparable to UK and New Zealand and among states to Michigan or Oregon or Colorado in the US.

its 71% irrigation needs (CGWB, 2021), the groundwater resources experience demand pressures from major urban centers like Lucknow, Varanasi and the National Capital Region of Delhi (see Figure 4.2.2, top panel).

Factors affecting GWL dynamics, i.e., hydrological attributes (aquifer structure and surface water sources), climatic factors (mild and dry winters and hot summers) and groundwater demand (irrigation and urbanization), are found to vary substantially across UP. A multi-layered alluvial aquifer system and dense network of rivers and canals along with abundant annual rainfall of 784 milimeters (mm) together form one of the richest groundwater repositories in the world. However, there exists a west to east trend in GWLs - up to 30 meters (m) deep in the western districts while less than 3 m in the eastern districts. Such spatial trends in GWL could be explained (at least partly) due to the weather patterns rerpesented by cool and dry weather (464 mm average annual rainfall) in the northern and western parts of the state while hot and wet weather (1,245 mm average annual rainfall) in the southern and eastern parts of the state (Guhathakurta et al., 2020; Muñoz Sabater, 2019). See Figure 4.2.2, bottom left and bottom right panels).

GWL data are periodically sampled each year through a network of monitoring wells owned (separately) by the Central Groundwater Control Board (CGWB) and the UP Groundwater Department (UPGWD)[6]. GWL monitoring is synchronized with the Monsoon cycle and cropping seasons.[7] In this study we utilize GWL monitoring data during 2009-2019 made available by UPGWD. GWL data were scraped from PDF documents made available on the state web-portal *https://upgwd.gov.in* (UPGWD, 2020). The dataset provides a unique ID and geographic coordinates for each well along with groundwater depth from ground surface (measured in meters) recorded twice every calendar year.

Figure 4.2.2 shows a map and annual distribution of the wells monitored between

[6]The CGWB is the apex-body for groundwater management in India. State-level agencies like the UPGWD conduct independent monitoring each year, however CGWB and state-level agencies jointly publish a state of local groundwater development report every five years since 1995 (CGWB, 2017, 2019, 2021). Here resource assessments are conducted at the block-level according to the Groundwater Estimation Committee norms (CGWB, 2009, 2017; CGWB, 2019).

[7]Monsoon rainfalls occur between June and September each year and GWLs are typically monitored during May (pre-Monsoon) and August-September (post-Monsoon) periods. Further, the major cropping seasons are *Kharif* (wet season; June-October), *Rabi* (dry season; November-March) and *Zaid* (summer season; April-May).

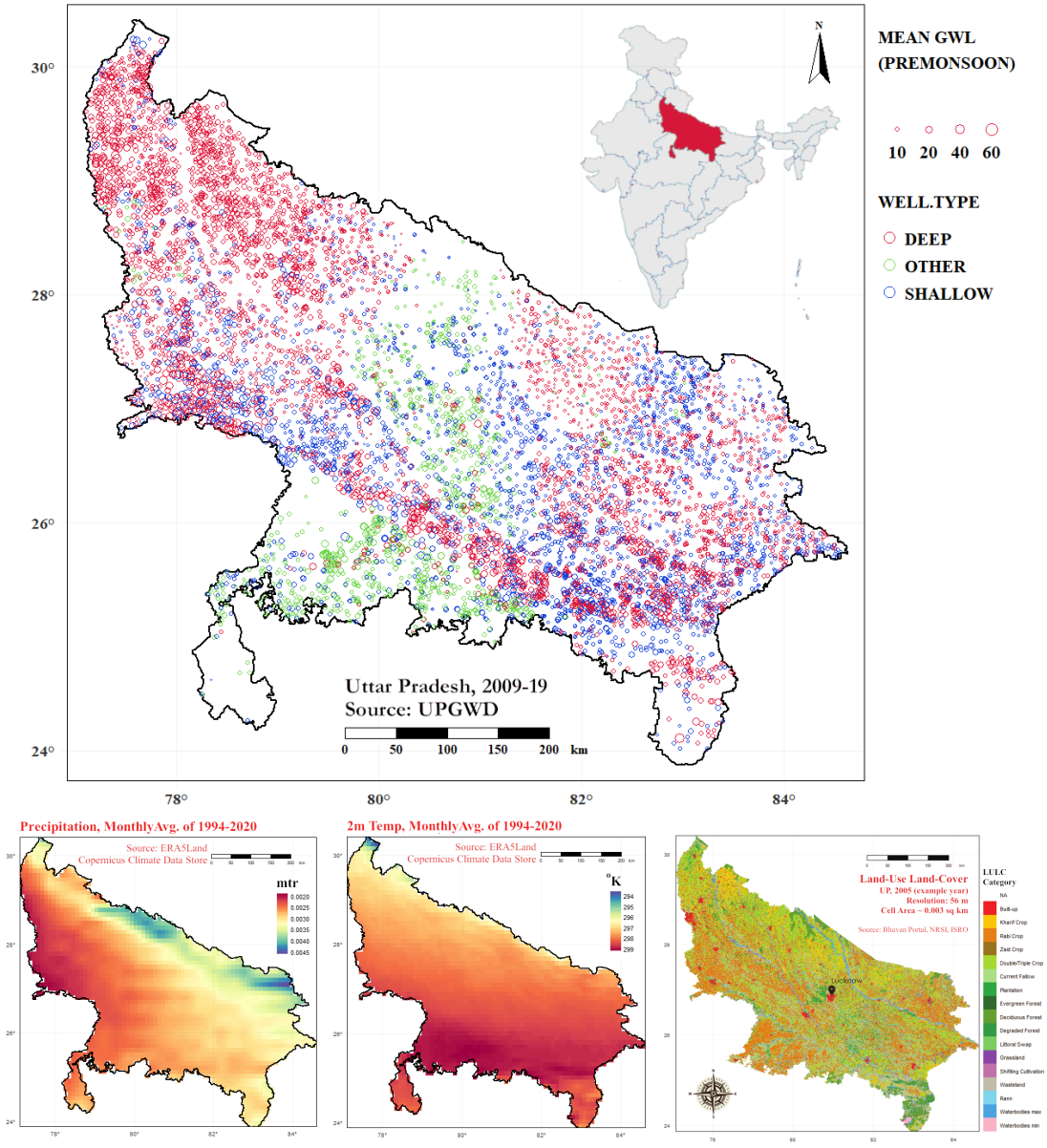


Figure 2.2: **Study Area, Monitoring Network, GWLs, Climate & Land-Use.** *Top panel*: Study area, Uttar Pradesh (UP), India, 2009-19 and UPGWD Monitoring network, wells shown as circles colored by type and sized by mean (of all years) pre-monsoon (May) GWLs. *Bottom left panel*: Monthly average precipitation between 1994-2020 in meters. *Bottom mid panel*: Monthly average temprature between 1994-2020 in degress Kelvin. *Bottom right panel*: Land-use Land-Cover categorical raster, 2005.

2009 and 2019 by UPGWD. Monitoring wells are characterized according to maximum depth, i.e., deep wells (average GWL depth $\approx 31.5 \pm 0.2$ m) and shallow wells (average GWL depth $\approx 13.4 \pm 0.2$ m). We utilize UPGWD dataset instead of the CGWB dataset because: (1) the UPGWD monitoring network comprises of 13,602 wells relative to only 1,376 CGWB wells during 2009-2019; and (ii) the UPGWD data provides textual labels identifying the cause of missing GWL values. The CGWB data does not provide any such information.

Missing GWL values were typically annotated as a blank, "NA" or a hyphen. Approximately 30% of the GWL records were missing during 2009-2019 and percentage missingness increased from 16.5% in 2009 to 47% in 2019 (see Figure 2.3). Only 13% wells provided complete GWL data in each season of each year during the study period. 11% were never monitored. 38.5% wells are monitored intermittently and 23% wells dropped out permanently in some year during the study period[8]. Missing data were spatially clustered with higher incidence of missingness in the central and northeastern parts of UP (relative to western UP) (see Figure 2.3, bottom-right panel).

About a third of the missing GWL data were appended with textual labels such as "DRY", "CLOSED", "CHOKED" and so on, indicating the underlying reasons for missing GWL values. Figure 2.3 (top-right panel) shows the percentage occurence of the different causes of missingness. For simplicity, we have condensed all miscellaneous causes into one category called "OTHERS". We used a PDF decoder to convert the PDF scans into comma-separated textual format and applied text processing commands in **R** to clean up and extract these labels.[9] The textual label "DRY" was interpreted to imply that the water level dropped below the corresponding monitoring well's maximum floor depth. On the other hand, labels such as "CLOSED" and "CHOKED" indicated, respectively, that the corresponding well was closed by the agency in a planned manner and that it got choked by a stray object (UPGWD, 2020). Figure 2.3 (top-right panel) plots the frequency of the most frequently occurring labels.

[8]Dropout is significantly higher in shallow wells as compared with the deep wells, indicating a shift in monitoring from shallow wells to deep wells, which is also corroborated by government reports (CGWB, 2019c).

[9]The underlying reasons for missing data are typically not reported in data publications made available through a standardized nationwide data dissemination portal known as the Indian Water Resource Information System (India-WRIS, 2022). Furthermore, UPGWD data are only made available for two years, 2018 and 2019, on the India-WRIS portal.

Dry wells are the most frequent *known* cause of missing GWL values accounting for about 3% of missing values; 20% are missing without a known cause and these likely include unreported dry wells. The incidence of dry wells increased from 0.2% (22 wells) in 2009 to 10% (809 wells) in 2019 (see Figure 2.3, bottom-left panel). Dry wells occur as clusters proximate to the urban centers and in eastern UP. Missing GWL records were more frequent among 'dry' shallow wells relative to 'dry' deep wells, and missing wells were more frequently observed to be 'dry' during the pre-Monsoon season relative to the post-Monsoon season.[10]

The The CGWB spatially aggregates well-level pre-Monsoon and post-Monsoon GWL observations to the level of an administrative block (CGWB, 2017; CGWB, 2019). The fact that spatially aggregated entities are important policy objects raises a natural inquiry as to how groundwater resource assessments might be impacted due to well-dryness related non-random missingness. A block-level map showing estimated groundwater stock appeared in the national compilation on dynamic groundwater resources of 2020 (CGWB, 2021). The compilation report does not address missingness issues leading to the possibility that a bias was introduced in stock estimation due to the bias in the spatial block-level means. The block-level stock estimate is used in the classification of each block in the country into one of four categories[11]. The CGWB notifies regulation such as restrictions on the number of groundwater abstraction and rainwater harvesting structures in an area, or the refinance of loans for tubewells based on the block classification so biases in the stock estimates may have real-world consequences (UPGWD, 2019; CGWB, 2019a; IRMED, n.d.). We analytically characterize the bias in spatial means due to missing GWLs data and provide empirical estimates of the size of the bias for all blocks in UP.

Besides this we address conditional GWL mean estimations which are also policy-relevant metrics. In addition to the data from groundwater monitoring agencies like the CGWB, the Minor Irrigation Census (MIC) is an important source of data for groundwater studies focused on India (MI Census, 2013b). Unlike the CGWB, the MIC does not provide the original well-level data at all publishing only the spatial village-level

[10]The gradual replacement of shallow wells by deep wells in the UPGWD monitoring network could be a strategy to reduce well-dryness related missingness, however we do not find a mention of such a strategy in the policy reports or data publciations.

[11]Safe, semi-critical, critical and over-exploited

means of GWL level[12]. Bhattarai *et. al* explain the impact of warming temperature on GWL declines using well-level GWL data (Bhattarai et al., 2023). Gupta utilizes village-level GWL data to estimate the change in conditional spatial mean of GWLs due to the causal impact of electricity pricing reform using a difference-in-difference (DID) regression model[13] (Gupta, 2021). Sekhri uses village-level data as regressors to study the impact of access to groundwater on poverty in villages in UP (Sekhri, 2014).

The truncation bias due to non-random missingness on statistical inference of means and conditional means is well studied in the statistics literature (Little & Rubin, 2019; Tang & Ju, 2018). We present a novel framework for the delineation of the source of the bias (dry wells), rigorous analytical proof for the non-random nature of missingness, demonstration of the bias in an experimental setting, identification the factors that contribute to missing data, the prediction of the bias and provision of confidence intervals as well as the interpretation of the impact of the bias *specifically* for groundwater management policy. Our framework is based on an integrated approach that involves a combination of analytical, empirical and simulations-based approaches. To our best knowledge, this is the first adaptation of non-random missingness theory to the empirical conundrum posed by dry wells in the context of groundwater resource management.

2.3 Materials and Methods

We posit GWL to be a positive, real-valued random variable $G \in \mathbb{R}^+$ such that $G \sim f_G(\theta)$ where $f_G(.)$ represents the probability density function (pdf) of G. In this section we outline our methodology to evaluate how missing GWL values *due to dry wells* impact sample-based θ estimator(s) in univariate (e.g., unconditional mean estimate over a spatial domain like administrative blocks) and multivariate (e.g., conditional mean estimate via regression analysis) settings. Our primary unit of analysis is a monitoring well i where GWLs are measured for time-periods $t = 1, ..., T$. Considering n monitoring wells in a region we organize GWL data as a column-vector $\vec{\boldsymbol{g}}_{nT\times 1}$, which contains monitoring instances $g_{i,t},\ i = 1, ..., n$ and $t = 1, ..., T$.

[12]The MIC reports "Average Ground Water Level (in meters)" in a village. See the "Village Schedule" of the 5th MI Census (MI Census, 2013a)

[13]The study area is Punjab, located close to UP within the same alluvial aquifer system.

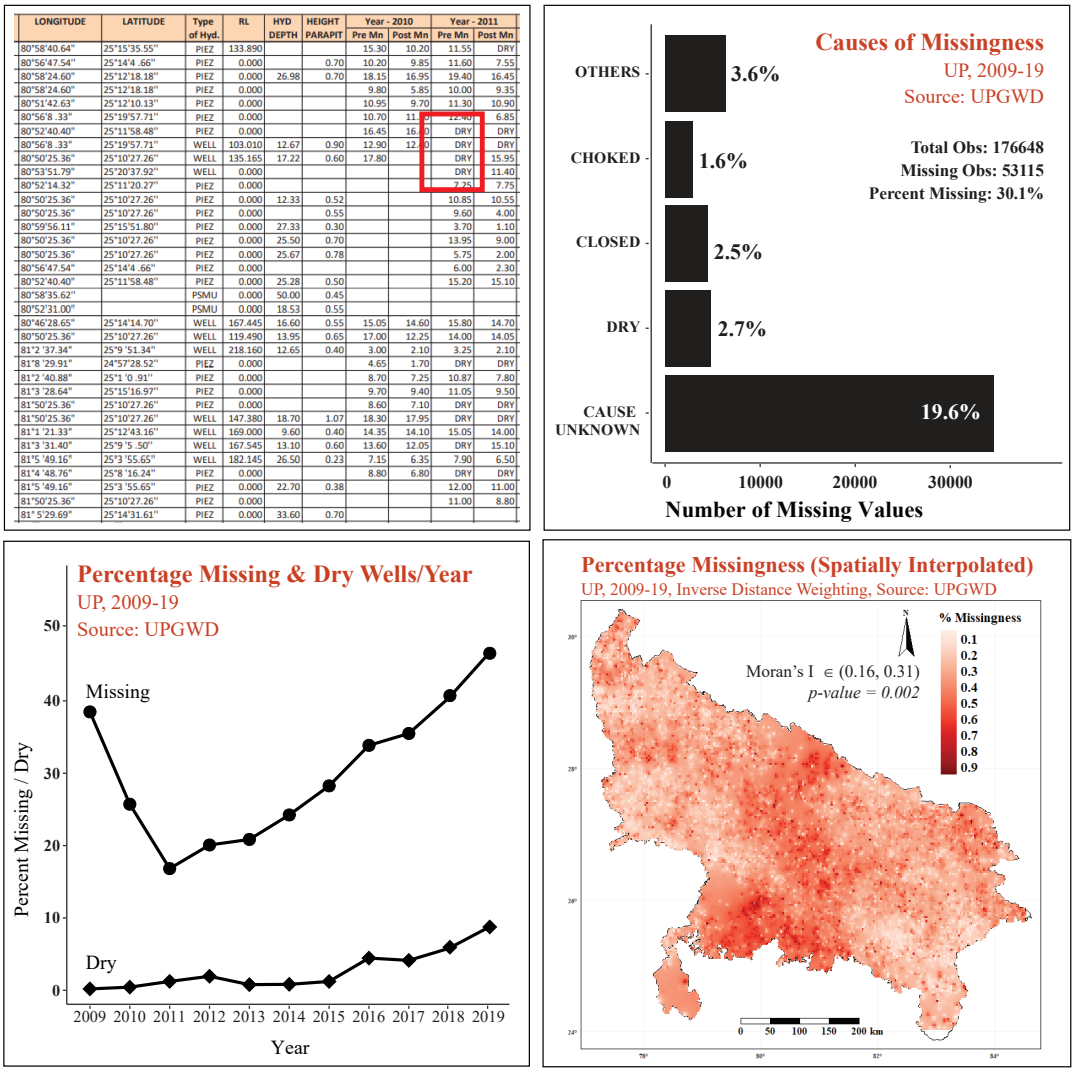


Figure 2.3: **Case Study of Missingness, UP 2009-19**. *Top-left panel*: Raw data, screenshot of a portion of the PDF file scraped from UPGWD, note that some missing observations are labeled "DRY" while others are blank, i.e, the cause of missingness is unknown. *Top-right panel*: Percentages of different causes of missingness. *Bottom-left panel*: Percentage of wells with missing values and percentage of wells dry by year. *Bottom-right panel*: Percentage missingness by well, spatially interpolated using inverse-distance weighting. Source: UPGWD.

With the knowledge that $g_{i,t} \in \mathbb{R}^+ \ \forall \ i,t$ we focus on missing GWL instances where $g_{i,t}$ is greater than maximum floor depth, denoted as $\bar{g}$ (see Figure 2.1), and $g_{i,t} > \bar{g}$ implies that well i was "dry" in year t[14]. We posit GWL missingness as a binary-valued random variable $M \in \{0,1\}$ such that $M \sim f_M(\phi)$ denotes the pdf of M. An indicator variable $m_{i,t}$ records missing observations where $m_{i,t} = 1$ denotes a missing observation (i.e., $g_{i,t} > \bar{g}$), and $m_{i,t} = 0$ denotes a valid, non-missing GWL record (i.e., $g_{i,t} \leq \bar{g}$). Similar to GWL data the sample-based missingness records are organized a column-vector $\vec{\boldsymbol{m}}_{nT\times 1}$, comprising of $m_{i,t},\ i = 1,...,n$ and $t = 1,...,T$. We append GWL observations with missingness records to organize observed GWL data in a sub-vector $\vec{\boldsymbol{g}}_{(0)}$ and missing GWL data in another sub-vector $\vec{\boldsymbol{g}}_{(1)}$ such that $\vec{\boldsymbol{g}}^T = (\vec{\boldsymbol{g}}^T_{(0)}, \vec{\boldsymbol{g}}^T_{(1)})$ (Rubin, 1976; Little & Rubin, 2019)[15].

To clarify, consider a cross-sectional setting where d out of n wells are dry in monitoring year t. Then the missingness-vector $\vec{\boldsymbol{m}}_{n\times 1}$ will induce a binary partition on $\vec{\boldsymbol{g}}_{n\times 1}$[16] into a $(n-d)\times 1$ sub-vector $\vec{\boldsymbol{g}}_{(0)}$ comprising of *observed* GWL values, i.e., $g_i \in (0, \bar{g}]$, and a $d \times 1$ sub-vector $\vec{\boldsymbol{g}}_{(1)}$ comprising of *missing* GWL values, i.e., g_i equals some unknown value g^* such that $g^* \in (\bar{g}, \infty)$.

Now given that GWLs remain partly observed in a region due to dry wells we include k covariates of GWL, denoted as $\boldsymbol{X}_{nT\times k}$, in order to further characterize the impact of missingness on sample-based inference. Specifically, we seek to ensure that for each period t covariates in $\boldsymbol{X}_{n\times k}$ are "fully observed" corresponding to sub-vectors $\vec{\boldsymbol{g}}_{(0)}$ and $\vec{\boldsymbol{g}}_{(1)}$. Hence, for the cross-sectional setting, the "full data" are described as a matrix $\{(\vec{\boldsymbol{g}}^T_{(0)}, \vec{\boldsymbol{g}}^T_{(1)})^T, \vec{\boldsymbol{m}}_{n\times 1}, \boldsymbol{X}_{n\times k}\}$ of size $n \times (k+2)$ while the "observed data" are a subset of the full data matrix denoted by $\{\vec{\boldsymbol{g}}_{(0)}, \vec{\boldsymbol{m}}_{n-d\times 1}, \boldsymbol{X}_{n-d\times k}\}$.

We restrict our analysis to "univariate" missingness whereby missing values arise only in the GWL variable while vectors **m** and **X** are fully observed. We ensure that

[14] Missingness in GWL data may be due to multiple causes (see Section 2.2), however the one that interests our study is missingness due to dry wells.

[15] Note that this is the simple case of "univariate" missingness since G is scalar-valued. G may also be vector-valued; for example, if we consider the pre-monsoon and post-monsoon GWL observations as a single data point then $G \in \mathbb{R}^+ \times \mathbb{R}^+$. In this case, the induced multivariate missingness variable, $M \in \{0,1\}^2$, is also vector-valued. This can be generalized to any dimension. In this paper, we do not consider dimensions greater than two although there are cases when missingness in up to 4 GWL observations can be considered together, as is the case with CGWB GWL data where there are 4 observations in each monitoring year.

[16] When t is fixed we will drop the time-subscript to simplify exposition. That is, in period t, $\vec{\boldsymbol{g}}_i$ of size $n \times 1$ that comprises of monitoring instances $g_i,\ i = 1,...,n$.

there is no missingness in $\mathbf{X}$ by constructing a set of well-level spatially-delineated high-resolution covariates that drive GWL dynamics, including weather variables made available by the Indian Meteorological Department and remotely-sensed land use land cover change variables made available by the Indian Space Research Organisation. To construct these covariates a "spatio-temporal neighborhood" for each well i in year t is defined as $N_i(r,l) = \{ (x_i', y_i', t') : s.t. |(x_i', y_i') - (r_i, g_i)| \leq r \; and \; t' = t - l \}$ where (r_i, g_i) denote at well i's location coordinates, r is the Euclidean distance between well i and its spatial neighbor with coordinates (x_i', y_i'), and t' is l years prior to the monitoring year t. We set $r = 1$ kilometer and $l = 1$.

Our covariates include total precipitation, average summer temperature, percentage area under multiple-cropping[17], percentage area under dry-season cropping[18], percentage area under urban settlement, groundwater runoff, distance to nearest river, distance to nearest highway and the number of human habitations within 5 km, . Hence, $k = 9$. Illustratively, total precipitation in spatio-temporal neighborhood $N_i(1,1)$ refers to total Monsoon rainfall received within a radius of 1 km around a well i in the year $t-1$. We provide details of data construction process in an Appendix 2.D. Table 2.4 provides detailed descriptions.

2.3.1 Missingness Mechanisms in Groundwater Data

Missingness mechanisms characterize how, if at all, missing data observations ($\vec{\boldsymbol{m}}_{nT\times 1}$) may exhbit linkage(s) with corresponding GWL observations ($\vec{\boldsymbol{g}}_{nT\times 1}$) and relevant covariates ($\boldsymbol{X}_{nT\times 5}$). This concept was formalized by (Rubin, 1976; Mealli & Rubin, 2015; Little & Rubin, 2019) in the form of a conditional pdf of M such that $M|(G,\mathbf{X}) \sim f_M(\phi|(G,\mathbf{X}))$, given the fixed covariates and the parameter ϕ of the distribution for missingness. Below we provide a brief description of three missingness mechanisms that are based on the the conditional density of missingness in GWL observations (Enders, 2010; Rubin, 1976).

[17] Multiple cropping is the process of cultivating the same piece of land more than once a year
[18] Indicating higher dependence on irrigation by wells

Missing Completely at Random (MCAR)

GWL data are said to be *missing completely at random* (MCAR) if and only if the incidence of missing values is independent of GWL itself (i.e., $g_{i,t}$) or *any* of its covariates. Mathematically, we can express MCAR as follows:

$$MCAR : f_M(\phi|(G, \boldsymbol{X})) \;=\; f_M(\phi) \tag{2.1}$$

The above equation states that the pdf of M is independent of the *entire* data, i.e, $\vec{\boldsymbol{g}}$ and $\mathbf{X}_{nT\times 5}$, meaning that the process that causes missing GWL values is fully exogenous to the processes that generates GWL data and its covariates. An important implication is that MCAR missingness is *ignorable* because the "observed" data are a random sample representation of the "full" GWL data[19], and deleting MCAR data will not affect unbiasedness of sample inference adversely (Rubin, 1976; Enders, 2010). While unbiasedness remains unaffected, deleting MCAR data will reduce the sample size; therefore, it generally affects variance of the estimates of the parameter. In other words, the standard error is likely to be larger.

Missing at Random (MAR)

GWL data are said to be *missing at random* (MAR) if the incidence of missingness in GWL data is independent of sub-vector $\vec{\boldsymbol{g}}_{(1)}$, i.e., the unobserved GWLs, *conditional on* $\mathbf{X}$. Mathematically, we can express MAR as follows:

$$MAR : f_M(\phi|(\vec{\boldsymbol{g}}_{(0)}, \vec{\boldsymbol{g}}_{(1)}, \boldsymbol{X})) \;=\; f_M(\phi|(\vec{\boldsymbol{g}}_0, \boldsymbol{X})) \tag{2.2}$$

The MAR class of missingness mechanisms may appear to imply that the incidence of missingness is independent of GWL data. However, Enders (2010), pp. 6 provided the clarification that MAR *actually means* that there is a systematic relationship between incidence of missingness and the observed GWL data. On the other hand, the probability of incidence of missingness is *independent* of missing GWL data conditional on

[19]For example, randomly deleting observations from "full" GWL data at the flip of a coin is an MCAR mechanism.

the observed covariates ($\mathbf{X}$) (Briggs et al., 2003). Note that MAR is also *ignorable* but in practice, it is difficult to formally test for MAR, which would require evaluating whether missingness is independent of the unobserved GWL data $\vec{\boldsymbol{g}}_{(1)}$.

Missing Not at Random (MNAR)

GWL data are *missing not at random* (MNAR) if the incidence of missingness is partially explained by the missing or *unobserved* values of GWL itself. In the case of MNAR, the conditional probability distribution of M depends on the *missing* values $\vec{\boldsymbol{g}}_{(1)}$, even after considering the covariates. Therefore, we can write:

$$MNAR: \ f_M(\phi|(\vec{\boldsymbol{g}}_{(0)}, \vec{\boldsymbol{g}}_{(1)}, \boldsymbol{X})) \neq f_M(\phi|(\vec{\boldsymbol{g}}_0, \boldsymbol{X})) \tag{2.3}$$

MNAR is considered *non-ignorable* meaning that naively deleting MNAR data will alter the sample-based inference on θ; the parameter of the distribution of GWLs data (Rubin, 1976). Statistical tests for MNAR mechanisms requires the knowledge of missing values. However, we can argue that MNAR mechanism stems from the prior knowledge of data generating process of $\mathbf{m}_{it}$.

2.3.2 Non-random Missingness: Ignorability of Missing Data due to Dry Wells

We investigate ignorability of missing data due to dry wells when performing statistical inference on unconditional and conditional spatial means using truncated GWL data. Specifically, we want to determine *whether* and *by how much* inference will change when missing data are ignored, i.e, deleted from the sample. Given a set of GWL observations of which some are missing due to dry wells and a set of fully observed covariates, we want to ask the following questions. *Is the missingness in the realized GWLs data non-random? Could missingness be random for some other realization of data? Will deleting missing data from the sample bias the estimates of unconditional and conditional spatial means? If yes, what will be the extent of this bias?* We use a combination of **empirical**, **conceptual** and **simulations-based** approaches to answer

these questions.

Our goal in this section is to provide a conceptual argument for why the missingness that occurs due to dry wells is always non-ignorable, i.e, to show first that missingness depends on the missing values themselves and second that ignoring or deleting missing observations due to dry wells will incur estimation bias. According to Rubin missing data are theoretically not ignorable if missingness is MNAR. When investigating missingness mechanisms for non-experimental data in practice, it is generally not possible to exclusively categorize the type of missingness into one category. Graham (2009) states that the MCAR, MAR, and MNAR missingness mechanisms should not be thought of as mutually exclusive categories of missingness. The best way to think of all missing data is as a continuum between MAR and MNAR. If we take the example of GWLs data, it may theoretically be possible to find a set of covariates that when observed with enough precision can explain GWL dynamics leaving only random residual variation but in practice such a set of covariates cannot be obtained. GWL dynamics in particular are difficult to explain due to the occurrence of the resource under the ground and the unavailability[20] of aquifer maps that reveal the structure of the geological water bearing formations. Some variation in GWLs is thus always present in the unobserved term. We need to answer the question of whether the unexplained part of GWLs is big enough for the truncation bias to matter to any practical extent (Graham, 2009, p.567).

Missingness is MNAR if the conditional density of missingness depends on the missing data itself even after considering the covariates, i.e $f_M(\phi \mid \vec{\boldsymbol{g}}_{(0)}, \vec{\boldsymbol{g}}_{(1)}, \boldsymbol{X}) \neq f_M(\phi \mid \vec{\boldsymbol{g}}_{(0)}, \boldsymbol{X})$. Mealli and Rubin (2015) provide the condition for missingness to be considered MNAR which we have adapted to the case of GWLs in the following definition.

Definition 2.3.2: In a sample $\vec{\boldsymbol{g}}_{n\times 1}$ containing n GWL observations (without associated covariates) where the observed portion of the sample is $\vec{\boldsymbol{g}}_{(0)}$ and the missing portion is $\vec{\boldsymbol{g}}_{(1)}$, the missing data are MNAR if $f_M(\phi| \vec{\boldsymbol{g}}_{(0)}, \vec{\boldsymbol{g}}_{(1)}, \vec{\boldsymbol{m}}_{n\times 1})$ does not equal $f_M(\phi| \vec{\boldsymbol{g}}_{(0)}, \vec{\boldsymbol{g}}'_{(1)}, \vec{\boldsymbol{m}}_{n\times 1})$ for some ϕ and for some $\vec{\boldsymbol{g}}_{(1)} \neq \vec{\boldsymbol{g}}'_{(1)}$. Note that the inequality must hold even after accounting for the observed data $\vec{\boldsymbol{g}}_{(0)}$.

Based on definition 2.3.2, we make the following proposition.

[20] For the Indian case, at least.

Suppose we have a sample $\vec{\boldsymbol{g}}_{n\times 1}$ containing $\boldsymbol{n}$ observations of GWL measured via $\boldsymbol{n}$ monitoring wells at different locations in space at a fixed time where the observed GWLs are in $\vec{\boldsymbol{g}}_{(0)}$ and the missing GWLs are in $\vec{\boldsymbol{g}}_{(1)}$ so that $\vec{\boldsymbol{g}}_{n\times 1} \equiv \{(\vec{\boldsymbol{g}}_{(0)}^T, \vec{\boldsymbol{g}}_{(1)}^T)^T\}$. Let the probability that any given well is dry be denoted by $\boldsymbol{p}_{dry}$. If $0 < \boldsymbol{p}_{dry} \leq 1$ then missingness is MNAR.

Let the realized missingness be represented by the vector $\vec{\boldsymbol{m}}_{n\times 1}$. Definition 2.3.2 states that the missingness is MNAR if probability $f_M(\phi|\, \vec{\boldsymbol{g}}_{(0)}, \vec{\boldsymbol{g}}_{(1)}, \vec{\boldsymbol{m}}_{n\times 1})$ of observing $\vec{\boldsymbol{m}}_{n\times 1}$ when the realized value of the missing GWLs is $\vec{\boldsymbol{g}}_{(1)}$ does not equal the probability $f_M(\phi|\, \vec{\boldsymbol{g}}_{(0)}, \vec{\boldsymbol{g}}'_{(1)}, \vec{\boldsymbol{m}}_{n\times 1})$ when the realized values of missing GWLs are $\vec{\boldsymbol{g}}'_{(1)}$ for some $\vec{\boldsymbol{g}}_{(1)} \neq \vec{\boldsymbol{g}}'_{(1)}$.

Let us first define some basic probabilities.

Probability of a well being dry: If the GWL at well i, g_i, falls below the maximum depth $\bar{g}$ then the well i is said to be dry. Then the probability that a well is dry is:

$$\boldsymbol{p}_{dry} = f_G(G > \bar{g}) = 1 - f_G(\bar{g}) \tag{2.4}$$

Probability of a well not being dry: If the GWL at well i, g_i, is less than the maximum depth $\bar{g}$ then the well i is not dry. Thus the probability that a well is not dry is:

$$p_{notdry} = f_G(G \leq \bar{g}) = f_G(\bar{g}) \tag{2.5}$$

Probability of observing the GWL at a dry well: At a dry well, the GWL falls below the maximum depth so it cannot be observed and a missing value is certain to be recorded, i.e $p(M = 1 \mid G > \bar{g}) = 1$ and $p(M = 0 \mid G > \bar{g}) = 0$.

$$p_{obs|dry} = p(M = 0 \mid G > \bar{g}) = 0 \tag{2.6}$$

$$p_{miss|dry} = p(M = 1 \mid G > \bar{g}) = 1 \tag{2.7}$$

Probability of observing the GWL at a well that is not dry: If the GWL at the monitoring well is above its maximum depth it can potentially be observed with some non-zero probability, say p. Even if GWL is less than the maximum depth, due to the presence of causes of missingness other than well dryness, a missing value may still be recorded with some finite non-zero probability, i.e, $0 < p(M = 1 \mid G <= \bar{g}) < 1$.

$$p_{obs|notdry} = p(M = 0 \mid G <= \bar{g}) = p \tag{2.8}$$

$$p_{miss|notdry} = p(M = 1 \mid G <= \bar{g}) = 1 - p \tag{2.9}$$

Consider a case in which there are no dry wells and suppose d values are missing (due to other causes) in a sample containing a total of n GWL observations. We thus obtain a missingness vector $\vec{\boldsymbol{m}}_{n\times 1}$ in which values for $i = 1, ..., n - d$ are 0s and for $i = n - d + 1, ..., n$ are 1s. Assuming that GWL and missingness realizations are independent, the probability of obtaining the pattern of missingness $\vec{\boldsymbol{m}}_{n\times 1}$ is:

$$p(M = \vec{\boldsymbol{m}}_{n\times 1} | \vec{\boldsymbol{g}}_{(0)}, \vec{\boldsymbol{g}}_{(1)}; \phi) = p^{(n-d)}(1 - p)^{(d)} f_G(\bar{g})^n \tag{2.10}$$

Now assume a counterfactual realization of missing values $\vec{\boldsymbol{g}}'_{(1)}$ in which well j was dry where $1 \leq j \leq n - d$, i.e GWL $g_j > \bar{g}$. Assuming that GWL and missingness realizations are independent, the probability of obtaining the pattern of missingness $\vec{\boldsymbol{m}}_{n\times 1}$ would then be:

$$p(M = \vec{\boldsymbol{m}}_{n\times 1} | \vec{\boldsymbol{g}}_{(0)}, \vec{\boldsymbol{g}}'_{(1)}; \phi) = p^{(n-d-1)}(1 - p)^{(d)} f_G(\bar{g})^{n-1} \times (0 \times (1 - f_G(\bar{g}))) = 0 \tag{2.11}$$

Clearly from equations 2.10 and 2.11, if there is a finite, non-zero probability that a well is dry i.e $0 < f_G(\bar{g}) \leq 1$, then $p(M = \vec{\boldsymbol{m}}_{n\times 1} | \vec{\boldsymbol{g}}_{(0)}, \vec{\boldsymbol{g}}_{(1)}; \phi) \neq p(M = \vec{\boldsymbol{m}}_{n\times 1} | \vec{\boldsymbol{g}}_{(0)}, \vec{\boldsymbol{g}}'_{(1)}; \phi)$.

By definition 2.3.2, missingness is thus MNAR.

This conceptual proof depends on the prior knowledge of the probability of a well

being dry in a particular monitoring network. In theory, as long as wells have a finite maximum floor depth $\bar{g}$, there is always a finite non-zero probability that a well is dry but this probability may be close to zero in some cases where $\bar{g}$ is very large compared to the mean GWL. In general, we cannot assume for any GWL data that the maximum floor depth or the presence of dry wells is known apriori. We thus need to employ an empirical approach to characterize the missingness mechanism from observed GWLs data and the associated observed pattern of missingness[21]. To study missingness for any generally available GWLs data, we propose an empirical approach based on existing **statistical tests** for missingness and a novel method for predicting dry wells based on **logistic regression**. The approach consists of the following steps:

1. Assume that a finite, non-zero number of missing values have occurred due to dry wells
2. Select and construct a set of covariates that influence GWLs and thus will necessarily drive the dryness of wells
3. Use available statistical tests to rule out MCAR (completely random missingness) as a preliminary step. Note that these tests rely on the availability of covariates of missingness
4. Run a logistic regression to estimate the probability of missingess using the set of constructed covariates
5. Interpret the regression coefficients to see if they are consistent with (or contradict) the hypothesis that dryness is a major underlying cause of missingness
6. Use the estimated probabilities to predict which missing observations occurred due to dry wells.
7. Use the above predictions to perform truncation bias estimation for unconditional spatial means and conditional means of GWL. Note the truncation bias is the bias that results from listwise deletion of missing data that occurred due to dry wells.

Hora *et. al* (2019) postulate that the truncation of missing values due to dry wells leads to survivor bias. Survivor (or survivorship) bias refers to statistical estimation bias that results when individuals that *drop out* of the sample over time are systematically ignored and analysis is restricted only to those individuals that "survived" some elimination process like death or closure. Survivor bias may lead to erroneous inference

[21] In this paper, we are also able to exploit a unique dataset that provides confirmatory evidence that dry wells caused close to 3% of the missing values but we cannot assume in general that the causes of missing data are known.

sometimes called being "fooled by the winners" (Lockwood, 2021). Survivor bias is well studied in health economics where patients leave survey studies due to sickness or death (Heiss, 2011). In financial studies, it can occur when a study ignores companies that close during the study period focusing only on surviving companies (Brown et al., 1992).

In the case of groundwater monitoring wells, a well that goes dry in one monitoring period does not necessarily lead to dropout because it may fill up with water in one of the subsequent periods due to, for example, a good rainfall year. Dry wells may still lead to dropout sometimes and at other times manifest as a pattern of intermittent missingness. We find evidence that dry wells are associated with an intermittent missingness pattern far more often (82% of the time) as compared to dropout (12% of the time). We conclude that sample selection due to dry wells can potentially cause survivor bias but is different from the classic case of the survivor bias problem as previously indicated (Hora et al., 2019).

Another source of bias in spatial aggregation of GWLs can occur due to the modified aerial unit problem or MAUP. This refers to a statistical bias that occurs due to aggregation of point GWL measurement at wells at the level of spatial units like blocks. Changing the aggregation boundary may dramatically change the results. MAUP is especially problematic when the aggregation units are "arbitrary" and do not accord physically with the underlying data generating process. In India, the recommended spatial unit for groundwater resource assessment is the watershed which is the natural hydrogeological sub-division within an aquifer or basin but at present, resource assessment continues to be done along administrative boundaries (CGWB, 2017).

The issue of missing data is routinely encountered in economics, especially in labor and health economics. While naïve listwise deletion was often used as in Burkhauser (1995), more recent studies paid relatively greater attention to the issue (Burkhauser et al., 1995). Seah studied the effects of immigrant teachers on student's academic outcomes using data in which some observations of the teacher's native language were missing (Seah, 2018). The author argued that these observations were "randomly missing" by showing that the incidence of missingness was not correlated with other fully observed teacher or student characteristics. Statistically, this was shown by running

a set of regressions in which the fully observed variables were separately regressed on an indicator of whether or not the teacher's native language variable had a missing response or not. This is equivalent to the Little's test for missingness completely at random (MCAR) which we formally describe in the next section. The author thus listwise deletes all observations where teacher's native language was missing based on the premise that the missingness is random meaning that the deletion is not theoretically expected to influence statistical inference.

On the other hand, in a study of the changes in the characteristics of American youth, the authors fail to rule out random missingness finding observed characteristics of individuals who exited from the sample to be significantly different from those who stayed (Altonji, Bharadwaj, & Lange, 2012). They employ a response weighting scheme in which they use a probit model to estimate the probability of response for each observation and then use the inverse estimated probability to "weigh up" each response. In other words, giving greater weight the data of respondent individuals who are similar to non-respondents; a process known as inverse probability weighting (IPW). They also apply a robustness check by running a set of specifications that do not involve the variable that contains missing values on the full sample as well as a reduced samples in which the individuals who had missing responses removed (Altonji et al., 2012). In health economics, Forder and Allan (2013) use random imputation of missing values and find that it does not change their results significantly. They also test the robustness of their model by treating all individuals in a certain category to be missing and find it does not alter their results either (Forder & Allan, 2013). Abbasi *et. al* (2023) study lead exposure in children and proceed with a "selection-on-observables" assumption (an alternative name for MAR). Under this assumption, they test the robustness of their model by predicting missing values using the observed (Abbasi et al., 2023). Overall, the studies we reviewed resort to listwise deletion under an MCAR assumption, use a weighting scheme to weigh up observed values and/or predict missing values under an MAR assumption, make speculative arguments for why a certain missingness pattern occurs and use a number of robustness checks to demonstrate that missingness does not impact statistical inference.

In this paper, we deal with non-random missingness founded on the general ana-

lytical framework for missing data studies proposed by Rubin (Rubin, 1976) and Little (Little, 1988) and a novel logistic regression that models the probability of missingness using a constructed geo-spatial dataset and then use the estimated probabilities to estimate the truncation bias due to listwise deletion of dry wells from the sample. Next, we describe Little's test to rule out completely random missingness in a realized set of data based on testable statistical propositions.

Statistical tests for missingness

There is no way to statistically test for MNAR or MAR since they require that missingness be independent of missing (unobserved) data (according to the criteria in equation 2.3) (Enders, 2010, p.6). To show that missingness is independent of missing data requires us to know the true values of missing GWL observations which is not possible in non-experimental situations. In the case of MCAR, we can show that missingness in data is *not* MCAR but we cannot show that it is MCAR. To show that missingness is not MCAR, it is sufficient to show that missingess is not independent of the observed data. Thus MCAR is the only missingness mechanism that yields testable propositions (Enders, 2010).

Little (1988) proposed a test of "homogeneity of means" to rule out MCAR as a missigness mechanism in observed data. We pose the null hypothesis that the set of wells (individuals) where GWL data are observed and those where it is missing are similar with respect to some (fully observed) covariates like rainfall and irrigation. For example, if we find that mean rainfall around wells with missing GWL was statistically significantly different from mean rainfall around wells where they were available, there is sufficient evidence to rule out MCAR since the significant difference in means suggests that missingness is possibly related to the variation in rainfall (Enders, 2010; Little, 1988).

Let's say the first covariate of GWL contained in column 1 of the covariate matrix denoted $\vec{\boldsymbol{x}}_1$ represents total monsoon rainfall in the spatial neighborhood of a well. We group rainfall observations by missingness in GWL. The mean rainfall around wells where GWLs were observed ($m_i = 0$) is denoted as $\bar{\boldsymbol{x}}_{1,m_i=0}$ and $\bar{\boldsymbol{x}}_{1,m_i=1}$ is the mean rainfall around wells where GWL was missing ($m_i = 1$). If these two means are statisti-

cally significantly different, the difference suggests an association between rainfall and missingness in GWL. We pose the null hypothesis that there is no difference in means:

$$H_0 : \bar{\boldsymbol{x}}_{1,m_i=0} = \bar{\boldsymbol{x}}_{1,m_i=1},\ H_1 : \bar{\boldsymbol{x}}_{1,m_i=0} \neq \bar{\boldsymbol{x}}_{1,m_i=1} \tag{2.12}$$

Statistically significant difference between the two grouped means $\bar{\boldsymbol{x}}_{m_i=0}$ and $\bar{\boldsymbol{x}}_{m_i=1}$ rules out MCAR. The tests of homogeneity of grouped means are repeated with each of the remaining covariates.

2.3.3 Logistic regression: identifying systemic variation in missingness

Groundwater monitoring agencies typically provide GWL data containing a non-zero proportion of observations recorded as missing indicated by "NAs" or blanks. The premise of this paper is that the sub-population of the missing values in any such dataset that is attributed to dry wells constitutes non-random missingness which if deleted from the sample, will incur the estimation biases. We assume for this paper that the sub-population of missing values that occurred due to causes other than dry wells occurred at random and thus can be removed from the sample without adverse effects on inference. Since we expect truncation biases to occur only due to dry-well-related missingness, we must have recourse to some strategy for separating out the sub-populations of missing values within the data that occurred due to dry wells from that which occurred due to other causes.

In general, we cannot assume that monitoring agencies provide any details regarding missing values other than "NA" markers. We have obtained access, for this work, to a unique GWL dataset in which roughly a third of the missing observations have been explicitly labeled with the underlying cause of missingness ("DRY", "CLOSED" and so on) whereas 20% of the missing data are unlabeled (or "NA"). We have reduced the set of data labels to 5 categories. These categories of labels and their corresponding interpretations[22] are given in Table 2.3.

[22]These are "best-guess" interpretations by the authors. The monitoring agency does not provide any explanation for the labels.

Data Label	Interpretation
AVAILABLE	GWL observation available, i.e, not missing.
DRY	GWL observation missing due to dry observation well
CLOSED	GWL observation missing, well closed by monitoring agency.
CHOKED	GWL observation missing due to choked observation well.
OTHER	GWL observation missing due to known miscellaneous causes.
NO LABEL	GWL observation missing, cause of missingness is unknown.

Table 2.3: Data Labels and Interpretations

We exploit our labeled data to predict the number of missing values that occurred due to dry wells using logistic regression. We first remove all missing data due to causes other than dry wells from the sample. According to our assumption, such removal does not affect inference. We then estimate the probability of missingness as indicated by the binary indicator variable M conditional on a set of carefully constructed variables that are known to influence the probability of well dryness. $M_{i,t}$ is the outcome variable that indicates if a well i had a missing value in year t($M_{i,t} = 1$ if missing and 0 otherwise). We regress the log-odds of the outcome $M_{i,t} = 1$ on set of k' covariates $\boldsymbol{Z}_{nT\times k'}$ as follows:

$$ln_e(\frac{p(M_{i,t}=1|\boldsymbol{Z})}{1-p(M_{i,t}=1|\boldsymbol{Z})}) = \boldsymbol{Z}_{i,t}\gamma + \nu_{i,t} \tag{2.13}$$

Where the indicator $M_{i,t} = 1$ if the observation at well i was missing either due to a dry dry well or an unknown cause. We then select a set of covariates $\boldsymbol{Z}_{nT\times k'}$ that are known to drive the process of drying up of wells. Note that with such a model construction, on the left hand side, we will have noise in the data because many of the missing values for which the cause is unknown will have occurred due to causes other than dry wells. This type of missingness will partially be explained by the error term of our model $\nu_{i,t}$. Since we have especially selected covariates on the right hand side that drive well dryness, we expect that they will explain the missing data due to dry wells and not those that occurred due to other causes. Our hypothesis is that the final predicted probabilities of missingness from this model will be higher for wells that were dry determined by the selective inclusion of relevant covariates.

The probability that any given well is observed to be dry, i.e the probability that the

water level in the well has fallen below its maximum depth, is influenced by the type of well (DEEP or SHALLOW) and a set of climatic, hydro-geological and human factors that influence GWL dynamics (USGS, 2019). The processes that cause water scarcity and deepen the groundwater level will tend to increase the probability of wells drying up. We consider the impacts of rainfall, temperature, irrigation water use, domestic water use, urban development, groundwater and surface water interactions, runoff and aquifer characteristics on the probability of well dryness. We regress the log-odds of the event that a pre-monsoon (May) GWL observation is missing at well i in year t conditional on the covariates $\boldsymbol{Z}$ denoted by $p(m_{i,t}^{prem} = 1|\boldsymbol{Z})$ on the covariates $\boldsymbol{Z_{nT\times k'}}$ ($k' = 10$, which includes one interaction term). Table 2.4 lists and describes the selected covariates. The covariates are constructed at the well-level by extracting data within circular spatial neighborhoods of each well with radii of either 1 km or 5 km. For any well in the monitoring network, the nearest well is on average 2.8 km away. A circle of radius 1 km, in the average case, should contain no other well and a circle of radius 5 km, at most 1 other well.

The main source of groundwater recharge in UP is the Monsoon rainfall and is the most important determinant of groundwater availability in the region (CGWB, 2019). The Monsoon rainfall in UP starts in June and ends in September. The probability that a well was dry in May of year t depends on the amount of rainfall that was received around the well in the Monsoon of year $t-1$. We thus take the total amount of precipitation received within a 5 km radius of each well from June to September in year $t-1$ as our rainfall variable $p_{i,t-1}$. The choice of the 5 km radius is dictated by the spatial resolution of our climatic data which is available on a grid of 9 km $\times$ 9 km. Hotter temperatures can cause wells to dry by increasing loss of water through evapo-transpiration. We take the average monthly air temperature at 2 m above the ground in the months from January to May in year t as our temperature variable $k_{i,t}$.

We also consider the impact of water "runoff." Only a portion of water from rainfall infiltrates the top soil layer and stays stored in the aquifer. Otherwise, it either drains away over the surface (surface runoff), or under the ground (sub-surface runoff) and the sum of these two is simply called 'runoff'. The units of runoff are depth in metres. This is the depth the water would have if it were spread evenly over the area covered by

a grid cell of 9 km $\times$ 9 km. Deforestation, the draining of wetlands and the development of built-up settlements and paved surfaces can increase surface runoff and reduce groundwater recharge (Alley, 2009) which in turn can increase the probability of wells going dry through declining GWLs (Perrone & Jasechko, 2017b). We compute total runoff in the months from June to December in the year $t-1$ as our runoff variable $ro_{i,t-1}$.

We measure irrigation water use by the type of agricultural land-use that occured in the spatial neighbourhood of each well in the past year. Irrigation water withdrawals will be higher in volume around wells where multiple cropping or dry-season (*rabi* or Winter) cropping occurs as compared to wells located within wet-season (*kharif* or Monsoon) cropping occurs. We construct 2 land-use covariates that correlate with heavier irrigation water withdrawal. First, the percentage area under multiple cropping, denoted by $mc_{i,t-1}$; and second, the percentage area under dry season (*rabi* or winter) cropping, $dc_{i,t-1}$, within a 1 km radius of each well in the past year. The average size of operational agricultural landholdings in UP for the 2016 agricultural census was 0.73 hectare (0.0073 sq km) (Agricultural Census, 2016). This means, on average, our well neighborhood of radius 1 km will encompass 861 landholdings. We assume that irrigation within these holdings will generate sufficient irrigation water withdrawal to significantly influence GWL dynamics at the particular observation well.

We measure domestic water use around an observation well by the total population of habitations in a 5 km radius in the year 2016, denoted by $pop_{i,2016}$. Our assumption is that wells located in areas with greater population will experience a larger drawdown of GWL due to domestic water pumping in their neighborhood. We do not have spatially delineated data for population for each year in the study period. We assume that the spatial distribution of population in 2016 (roughly in the middle of our study period 2009-2019) to be representative of the population distribution for all years in the study period. The volumes of withdrawals are much lower for domestic water use as compared to irrigation water use so we expand the radius of measurement from 1 km (in the case of irrigation withdrawals) to 5 km for domestic withdrawals. We measure the effects of urban development by including the percentage area under built-up settlements in year $t-1$ around a 1 km radius of each well denoted by $b_{i,t-1}$.

It is well known that rivers and surface water bodies interact with groundwater systems due to seepage of water through river beds (Khan & Khan, 2019). The interactions of groundwater and surface water systems influence GWL dynamics and may have an effect on dryness of wells. We capture the effect of such interactions by constructing a variable that measures the distance of well i to the nearest a major river or tributary including major and minor tributaries (see appendix for details). We denote this variable by $dnrv_i$. We also consider the distance to the nearest major highway $dnrd_i$ as water demand maybe higher where there are highways.

We have well-level data from the monitoring agency on the "effective hydrological depth" (EHD). The EHD is related to the ability of the soil to store water. A greater EHD means that more water can be stored in the soil. Where EHD is greater, lesser surface runoff is produced (Sterk, 2021). This variable is negatively correlated to the runoff variable. The caveat is that for the EHD data, close to 60% of the observations are missing thereby significantly reducing the utility of the data for estimation. We thus omit this variable from our model specification.

A well goes dry when GWL falls below its maximum depth thus shallow wells are more likely to be dry compared to deep wells. We therefore report regression results separately for shallow and deep wells and also for the pooled dataset. Our baseline model specification for pooled data (both deep and shallow wells) is given by equation 2.14.

$$\begin{aligned} ln_e\left(\frac{p(M_{i,t}^{prem}=1|\boldsymbol{Z})}{1-p(M_{i,t}^{prem}=1|\boldsymbol{Z})}\right) = \gamma_0 + \gamma_1 p_{i,t-1} + \gamma_2 ro_{i,t-1} + \gamma_3 k_{i,t} + \gamma_4 mc_{i,t-1} + \\ \gamma_5 dc_{i,t-1} + \gamma_6 b_{i,t-1} + \gamma_7 pop_{i,2016} + \\ \gamma_8 dnrv_i + \gamma_9 dnrd_i + \gamma_{10}(p_{i,t-1} \times dc_{i,t-1}) + \nu_{i,t} \end{aligned} \tag{2.14}$$

Each estimated parameter $\hat{\gamma}_j$ is interpreted as the change in the log-odds of missingness when the corresponding covariate $\boldsymbol{Z}_j$ changes by one unit where the index $j = 0, 1, 2, ..., 9$ refers to the parameters. For example, $\hat{\gamma}_1$ is the expected increase in log-odds of missingness when total Monsoon rainfall in the past year $p_{i,t-1}$ increases by 1 unit.

Variable	Notation	Description
Missingness	$m_{i,t}$	Indicator, well i, pre-monsoon. 1 if missing, 0 if observed
Rainfall	$p_{i,t-1}$	Total Monsoon (Jun-Sep) rainfall, m (around* well i, year t-1)
Temperature	$k_{i,t}$	Mean monthly temp., Kelvin, Jan-May (around* well i, year t)
Runoff	$ro_{i,t-1}$	Total runoff (Jun-Sep), m (around* well i, year t-1)
Multiple-Crop	$mc_{i,t-1}$	Percent area multiple cropped (around** well i, year t-1)
Dry-Season Crop	$dc_{i,t}$	Percent area dry season cropped (around** well i, year t-1)
Built-up	$b_{i,t-1}$	Percentage area built-up (around** well i, year t-1)
Population	$pop_{i,2016}$	Total population (around* well i, year 2016)
Rivers	$dnrv_i$	Distance to nearest river, km (from well i)
Highways	$dnrd_i$	Distance to nearest hwy, km (from well i)
EHD	ehd_i	Effective Hydrological Depth (at well i)
Unobserved	$\nu_{i,t}$	Random unobserved characteristics (well i, year t)

Table 2.4: Variables in the logistic regression. *Within 5 km of well, **Within 1 km of well. See appendix section 2.D for details on data construction.

2.3.4 Predictive Model for Unlabeled Missing Data

While we have access to partially labeled missing data in this study, in the general case we expect missing values to appear without labels so we must employ a method for predicting dry wells. Our goal in this section is as follows; given an unlabeled missing GWL observation, use the logistic regression model proposed in the previous section to predict whether it occurred due to a dry observation well.

Say we have a sample containing n GWL observations in which d observations are missing due to dry wells. Since we have no labels, d is unknown. Our goal is to estimate d. We denote the estimate by $\hat{d}$. We apply a machine learning methodology to compute $\hat{d}$ in which we use the labeled data to train the model and then use the estimated coefficients to make predictions on the unlabeled data.

1. **Remove random missingness:** we first remove the labeled missing observations that occur due to causes other than dry wells. Based on our assumption, these are missing completely at random and their deletion from the sample has no effect on inference. Let's say that after this deletion, the number of observations we are left with is our n.

   Each of these n observations falls into one of the following categories:

(a) Observed (we label these NON-DRY, since observed wells cannot be dry)

(b) Missing and labeled DRY

(c) Missing and unlabeled

The set of observations falling into categories (a) and (b) constitute the "training" data that we will use to learn the values of the coefficients in our logistic regression model. The set of observations in category (c) is our out-of-sample data on which we will make predictions using the learnt coe fficients. Let's say that of a total of n well observations, say n_{lnd} are NON-DRY (i.e observed), n_{ld} are missing and labeled DRY and n_{ul} are missing and unlabeled. Our training data thus contains $n_{lnd} + n_{ld}$ observations and the out-of-sample data contains n_{ul} observations.

2. **Data split:** we split the training data using a 70-30 split in which we use 70% to train the model and 30% as testing data.

3. **Training:** we run the logistic regression model specified by 2.14 and obtain the vector of coefficient estimates $\hat{\gamma}$.

4. **Testing:** we then use the estimated coefficients to estimate the vector of $\hat{p}$, the probability of a well being dry, for each observation in the testing data (30%). We select a probability threshold p^* and if for an observation i, if $\hat{p} > p^*$, we assign it the label "DRY" and "NON-DRY" otherwise. We do this for three different values of p^*. See item 8.

5. **Confusion matrix:** we use the predictions from the previous step to build a confusion matrix which has the form shown in Table 2.5.

   We then compute two confusion metrics, the "precision" r_{tp} and what we term the "false negative prediction rate" r_{fn} as below:

$$r_{tp} = \frac{TP}{TP + FP} \tag{2.15}$$

$$r_{fn} = \frac{FN}{FN + TN} \tag{2.16}$$

   The precision is the fraction of predicted DRY wells that were actually DRY. The false negative prediction rate is the fraction of predicted NON-DRY wells that were actually DRY.

6. **Out of sample dry well prediction:** we use the estimated coefficients from the training phase to make predictions on our out-of-sample data containing unlabeled missing observations. This divides n_{ul} unlabeled missing observations into two sets of observations; $\hat{n}_{pd}$ observations that are predicted to be DRY and $\hat{n}_{pnd}$ that are predicted to be NON-DRY.

7. **Estimation of $\hat{d}$:** The estimated number of missing values due to dry wells $\hat{d}$ can now be estimated as the number of missing observations labeled DRY plus the number predicted DRY; $\hat{d} = n_{ld} + \hat{n}_{pd}$.

		Predicted Label		
		DRY	NON-DRY	Total
True Label	DRY	TP	FN	$TP + FN$
	NON-DRY	FP	TN	$FP + TN$
	Total	$TP + FP$	$FN + TN$	

Table 2.5: Structure of the confusion matrix. TP: True Positives, FN: False Negatives, FP: False Positives and TN: True Negatives.

To account for confusion, we compute an interval for the number of predicted DRY wells using the confusion metrics obtained in step 5. The lower limit of the interval is the product of the precision and the number of predicted DRY wells, i.e, $\hat{d}_{lower} = n_{ld} + r_{tp} \times \hat{n}_{pd}$. The upper limit is obtained by adding the product of the false negative prediction rate and the number of predicted NON-DRY wells, i.e, $\hat{d}_{upper} = n_{ld} + \hat{n}_{pd} + r_{fn} \times \hat{n}_{pnd}$. The idea is to correct for false positives and false negatives. False positives are wells that are predicted DRY even though they were not actually. These cause overestimation of DRY wells. False negatives on the other hand cause underestimation since these are predicted NON-DRY wells that were actually DRY. We posit that the estimated number of DRY wells, $\hat{d}$, lies within the interval $[\hat{d}_{lower}, \hat{d}_{upper}]$.

8. **Choice of p^*:** The choice of the probability threshold beyond which we predict a well to be DRY influences the final estimated $\hat{d}$. A lower value will cause overestimation whereas a higher value underestimation. We report our results for three different values corresponding to the first quartile, the mean and the third quartile of the estimated probability vector $\hat{p}$. For each value of p^*, we get a corresponding estimate of $\hat{d}$ as well as an interval $[\hat{d}_{lower}, \hat{d}_{upper}]$.

2.3.5 Bias Estimation for Real-World Data

Once we estimate the number of missing value due to dry wells $\hat{d}$ using the procedure described in the previous section, we can use the estimate to compute biases in statistical estimates of the unconditional spatial sample mean as well as the conditional mean of GWL. To account for sample selection biases in step 2 (Data Split), we use a Monte Carlo simulation in which we repeat the steps 2-8 for a pre-specified number of iterations N_{iter}. After each iteration, we obtain $\hat{d}$ and $[\hat{n}_{pd,lower}, \hat{n}_{pd,upper}]$. We use these estimates to compute truncation bias in the spatial sample mean GWL and the bias in the regression coefficients of the conditional mean GWL with the procedures described in the next two subsections.

Unconditional Spatial Mean Bias

Let $\hat{\mu}_{full}$ be the sample mean of the full data and $\hat{\mu}_{obs}$ be the sample mean of only the observed data. Let $\mu = E[G]$ be the population mean of the groundwater level so by definition of the sample mean, $E[\hat{\mu}_{full}] = \mu$. In the case of full data, the natural estimator for μ is the sample mean $\hat{\mu}_{full}$ given by equation 2.17.

$$\hat{\mu}_{full} = \frac{\sum_{i=1}^{n} g_i}{n} = \frac{\sum_{m_i=0} g_i + \sum_{m_i=1,dry} g_i + \sum_{m_i=1,oc} g_i}{n} \tag{2.17}$$

Where the last two summations are respectively the sums of the missing values due to dry wells and the missing values due to other causes. In general both types of missing observations (due to dry wells and due to other causes) are unknown so we cannot compute the full sample mean directly according to the formula in 2.17.

If we ignore the missing data as is the usual recourse, we can then compute the truncated mean $\hat{\mu}_{obs}$ using only the observed data. In a partially observed groundwater data sample G containing n observations at wells that each have the same maximum depth $\bar{g}$ there will be d observations ($0 < d < n$) missing due to dry wells and $\tilde{n}$ observations ($0 \leq \tilde{n} < n$) missing due to causes other than dry wells (which we denote "other-causes" or "oc"). The truncated sample mean is given by equation 2.18.

$$\hat{\mu}_{obs} = \frac{\sum_{i=1}^{n} (1 - m_i)\, g_i}{\sum_{i=1}^{n} (1 - m_i)} \tag{2.18}$$

We show in appendix 2.A that $\hat{\mu}_{obs}$ is a biased estimator of the mean μ and the magnitude of bias is directly proportional to the fraction of dry wells $\frac{d}{n}$. Since the full sample is generally unknown, we cannot estimate the actual size of the bias. We can however obtain a lower bound on the magnitude of the bias by using the definition of a dry well. We note that for if well observation i is missing due to a dry well, then missing value g_i must be greater than the maximum depth $\bar{g}$ by definition of dry wells. Thus, for all dry wells, $g_i > \bar{g}$. In other words, for dry wells the lower bound of g_i is $\bar{g}$.

We assume that:

1. The missing observations due to other causes are missing completely at random therefore they can be removed from the sample without any impact on inference. Thus dry wells will be the sole cause of missingness in the sub-sample.
2. For all missing values due to dry wells, GWL is just beneath the maximum depth of the well i.e we can make the substitution $g_i \approx \bar{g} \; \forall \; i$ s.t $m_i = 1$ (meaning for all missing values due to dry wells).

$$b_{min} = \hat{\mu}_{lb} - \hat{\mu}_{obs} = \frac{d(\bar{g} \; - \; \hat{\mu}_{obs})}{n} \tag{2.19}$$

Since $max(\hat{\mu}_{obs}) = \bar{g}$ (recall that $g_i \leq \bar{g}$ for observed values) we conclude that $b_{min} \geq 0$ which implies that $\hat{\mu}_{obs}$ is either less than or equal to $\hat{\mu}_{lb}$ (equality when there are no dry wells, $d = 0$). The difference between the sample means is directly related to d/n, the proportion of dry wells in the monitoring network. As the proportion of dry wells increases, the estimated sample mean becomes proportionally smaller than the full sample mean. Intuitively this makes sense since ignoring dry wells amounts to removing GWL values from the right tail of the sampling distribution.

We use Monte Carlo iteration to estimate the minimum bias for our real-world GWL data. In each Monte Carlo iteration, we follow the steps 1-7 described in 2.3.4 to estimate the total number of dry wells $\hat{d}$ and the lower $\hat{d}_{lower}$ and upper $\hat{d}_{upper}$ interval. Using these estimates from each iteration, we estimate the vector $(\hat{\mu}_{obs}, \hat{\mu}_{lb}, \hat{\mu}_{lb,lower}, \hat{\mu}_{lb,upper})$.

$$\hat{\mu}_{obs} = \frac{\sum\limits_{m_i=0} g_i}{n} \tag{2.20}$$

$$\hat{\mu}_{lb} = \frac{\hat{\mu}_{obs}(n - \hat{d}) + (\hat{d} \times \bar{g})}{n} \tag{2.21}$$

$$\hat{\mu}_{lb,lower} = \frac{\hat{\mu}_{obs}(n - \hat{d}_{lower}) + (\hat{d}_{lower} \times \bar{g})}{n} \tag{2.22}$$

$$\hat{\mu}_{lb,upper} = \frac{\hat{\mu}_{obs}(n - \hat{d}_{upper}) + (\hat{d}_{upper} \times \bar{g})}{n} \tag{2.23}$$

Using N_{iter} estimates obtained from the simulation, we estimate the expected values

and confidence intervals for each of the above quantities. We can then compute the biases b_{min}, $b_{min,lower}$ and $b_{min,higher}$.

Conditional Spatial Means Bias

A commonly used causal inference strategy is to estimate the conditional mean of GWL using as a function of fixed covariates and a vector of parameters using ordinary leasts squares (OLS). We are interested in the estimate of the regression coefficient(s) that measure the causal impact of change in a variable of interest on GWL. In this section, we describe our methodology to assess truncation bias that occurs in the coefficient estimates when missing values due to dry wells are deleted from the sample. Let's say we want to measure the causal impact of rainfall on GWL. Further, say we have a set of covariates $\boldsymbol{P}_{n\times k}$ that measure rainfall and GWL, as usual, is represented by the random variable G. We model the conditional mean of GWL as a linear function of rainfall, i.e, we posit a linear function of the form $E[G|\boldsymbol{P}_{n\times q}] = \boldsymbol{P}_{n\times q}\eta_{q\times 1}$.

We estimate the bias that occurs in the OLS estimates of η by the listwise deletion of missing observations due to dry wells. When d missing observations are removed from the full data sample, the corresponding covariates (rainfall) observations are also removed. Our full data is a set that contains n observations for GWL and each rainfall covariate denoted by $\{\vec{\boldsymbol{g}}_{(0)}, \vec{\boldsymbol{g}}_{(1)}, \vec{\boldsymbol{m}}_{n\times 1}, \boldsymbol{P}_{n\times q}\}$ where $\vec{\boldsymbol{g}}_{(0)}$ and $\boldsymbol{P}_{n\times q}$ are fully observed and $\vec{\boldsymbol{g}}_{(1)}$ are missing as indicated by $\vec{\boldsymbol{m}}_{n\times 1}$. The truncated data is $\{\vec{\boldsymbol{g}}_{(0)}, \vec{\boldsymbol{m}}_{n\times 1}, \boldsymbol{P}_{n-d\times 1}\}$ where $\boldsymbol{P}_{n-d\times q}$ is the truncated covariate matrix that remains after list-wise deletion of observations that correspond to missing values in $\vec{\boldsymbol{g}}_{n\times 1}$.

The conditional mean of G given the full data is expressed as $E[G \mid \vec{\boldsymbol{g}}_{(0)}, \vec{\boldsymbol{g}}_{(1)}, \vec{\boldsymbol{m}}_{n\times 1}, \vec{\boldsymbol{P}}_{n\times q}]$ and the conditional mean given the truncated data is expressed as $E[G \mid \vec{\boldsymbol{g}}_{(0)}, \vec{\boldsymbol{m}}_{n\times 1}, \vec{\boldsymbol{P}}_{n-d\times q}]$. The full data is unknown since we do not know the GWL for missing well observations. We employ Monte Carlo simulation to make our best assessment of bias according to the following methodology:

1. Estimate the linear regression with truncated data

$$E[G \mid \vec{\boldsymbol{g}}_{(0)}, \vec{\boldsymbol{m}}_{n\times 1}, \vec{\boldsymbol{P}}_{n-d\times q}] = R\eta + w \tag{2.24}$$

and obtain the estimates $\hat{\eta}_{trunc}$.

2. Using steps 2-4 from the predictive modeling procedure described in the section 2.3.4, we predict which among the unlabeled missing values occurred due to DRY wells.

3. For each well predicted (or labeled) DRY, make the substitution $g_i = \bar{g}$.

4. Estimate the linear regression after making the substitution and obtain the estimates $\hat{\eta}_{pred}$ which stands for "predicted" value of η based on DRY well prediction.

5. In each Monte Carlo iteration, estimate $\hat{\eta}_{trunc}$ and $\hat{\eta}_{pred}$ to get a set of N_{iter} estimates.

6. Report the expected value and confidence intervals of $\hat{\eta}_{trunc}$ and $\hat{\eta}_{pred}$ by averaging across the obtained estimates.

Choice of covariates: The base rainfall measurement is the total Monsoon precipitation received inside a 9 km by 9 km grid cell that is nearest to the well in the previous calendar year denoted by $p_{i,t-1}$. Rainfall is measured as the depth in meters of accumulated rainwater over the Monsoon period would have if spread out evenly over the surface of the entire grid cell. See appendix 4.4.2 for details. To capture non-linearity in the dependence of G on R, we construct three different covariates.

- Average rainfall at a well across the whole study period: $p_i = \dfrac{\sum_{t=0}^{T} p_{i,t}}{T}$
- Surplus rainfall in year $t-1$ with respect to the groundwater monitoring year t: $psur_{i,t-1} = max(p_{i,t-1} - p_i, 0)$
- Deficit rainfall (absolute value) in year $t-1$ with respect to the groundwater monitoring year t: $pdef_{i,t-1} = |min(p_{i,t-1} - p_i, 0)|$

Dependent Variable: The dependent variable is the pre-monsoon (May, prem) GWL at each well measured in meters below ground level (mbgl) denoted by $g_{i,t}^{prem}$.

Model Specification: Using the above dependent variable and covariates, we posit the following true model:

$$g_{i,t}^{prem} = \eta_0 + \eta_1 p_i + \eta_2 psur_{i,t-1} + \eta_3 pdef_{i,t-1} + v_{i,t} \tag{2.25}$$

where $v_{i,t}$ is a random error term. We estimate $\vec{\eta}$ in each Monte Carlo iteration to get $\hat{\eta}_{trunc}$ and $\hat{\eta}_{pred}$.

We also make the following simplifying assumptions that may or may not be true in practice but do not affect our results.

1. Assume that dry wells are the only cause of missingness in the data.
2. Assume that there is only one type of monitoring well in the monitoring so that $\bar{g}$ is a scalar constant that applies to all wells.

2.3.6 Bias Estimation for Synthetic Data - GWL and Rainfall

Estimating GWL response to Monsoon rainfall is of critical importance for groundwater budgeting as it helps to predict groundwater availability for human use given the amount of rainfall received in a particular cropping year (CGWB, 2017). This is especially the case for our study area where Monsoon rainfall is the largest source of groundwater recharge (CGWB, 2021; CGWB, 2021). The relationship between rainfall and GWL is a complex dynamic that depends on aquifer characteristics, geography, land-use patterns, soil quality, temperature and vegetation index. Numerous regional studies in hydrology have reported plausible linear relationships between rainfall and observed GWLs with R^2 ranging from 0.83 to 0.96 (Hussain, Wu, & Shih, 2022) while non-linear relationships have also been explored (Cobb & Harvey, 2019). As an initial foray in studying the impact of missing data, we restrict ourselves to the linear case. We consider a simple linear regression of the GWL G at a well i on the amount of Monsoon rainfall R received in a 1km radius of the well as specified by the equation 2.26.

$$G = R\beta + u \tag{2.26}$$

where $u \sim N(0, \sigma^2)$ is the error term. The parameter vector β is a 2×1 vector containing the constant of the regression β_0 and the slope of rainfall β_1. We want to estimate the parameters β from a sample of n observations $\{\boldsymbol{g}_{n\times 1}, \boldsymbol{r}_{n\times 1}\}$ where the $r_i's$ are fully observed but some of the $g_i's$ maybe missing indicated by the column vector of binary missingness indicators $\vec{\boldsymbol{m}}_{n\times 1}$. We assume that rainfall is exogenous to the

GWL and thus the regression given by equation 2.26 captures a directional causality from right to left quantified by the estimated value $\hat{\beta}_1$.

We also make the following simplifying assumptions that may or may not be true in practice but do not affect our results.

1. Assume that dry wells are the only cause of missingness in the data.
2. Assume that there is only one type of monitoring well in the monitoring so that $\bar{g}$ is a scalar constant that applies to all wells.

While the bias estimation method based on dry well prediction for real-world data (described in section 2.3.5) gives us the opportunity to examine the bias that may occur while studying the causal impact of rainfall on GWL, it is prone to confusion in the prediction step. The occurrence of type I and type II errors in the logistic regression model for dry well prediction will reduce confidence in the estimated bias. Further, it does not allow us to study estimation bias under different types of GWL and rainfall regimes.

We supplement our results with an illustrative Monte Carlo simulation study using synthetic data to factor out prediction error. We simulate the behaviour of the ordinary least squares (OLS) estimator $\hat{\beta}_1$, the coefficient of rainfall as missing GWL values due to dry wells are deleted from the sample along with the corresponding values for rainfall. We are interested in the relative performance of the estimator when the regression is performed with the full data versus the truncated regression with only the observed data.

We are also interested in the impact of the *existing* relationship between GWL and rainfall on the estimation bias. To study different GWL and rainfall relationships, we examine the behaviour of the OLS estimator as a function of the correlation ρ_{RG} between rainfall and GWL. Based on prior knowledge of the hydrologic cycle, we know that rainfall is typically contributes to groundwater recharge and areas receiving heavier rainfall have shallower (or lower) GWLs. Rainfall is thus inversely correlated with the GWL. We therefore expect the true value of the slope parameter of lagged rainfall β_1 and the correlation ρ_{RG} to be strictly less than zero. We simulate jointly distributed, correlated for both GWL and rainfall data under four different scenarios in which the

correlation is low-negative, high-negative, low-positive and high-positive. We detail the simulation procedure in appendix section 2.B.

We interpret our simulation results based on a preliminary analytical exercise. We show in Appendix 2.C that if $\bar{r}_{full}$ and $\bar{r}_{obs}$ are the sample means of the rainfall observations with the full and observed data respectively and similarly $\bar{g}_{full}$ and $\bar{g}_{obs}$ for GWLs and for observations $i = 1, ..., r*$ the GWLs are observed and for $i = r*+1, ..., N$ they are missing, then we have for the full data:

$$\hat{\beta}_{1,full} = \hat{\beta}_{1,obs}.k + \frac{(\bar{r}_{obs} - \bar{r}_{full})\bar{g}_{obs}}{\hat{\sigma}^2_{r,full}} + \frac{\sum_{i=r*+1}^{n} (r_i - \bar{r}_{full})g_i}{\hat{\sigma}^2_{r,full}} \tag{2.27}$$

where $\hat{\sigma}^2_{r,full}$ is the variance of rainfall in the full data and $\hat{\sigma}^2_{r,obs}$ is the variance of the observed data and $k = \frac{\hat{\sigma}^2_{r,obs}}{\hat{\sigma}^2_{r,full}}$. The first condition for equality between $\hat{\beta}_{1,full}$ and $\hat{\beta}_{1,obs}$ is $k = 1$ i.e there is no change in variance of rainfall by removing missing values. The second and third terms in equation 2.27 suggest that bias increases if the covariance between rainfall and GWL increases. We test these analytical assertions through our Monte Carlo simulation.

2.4 Results and Discussion

We report results from an illustrative case study of missingness for GWLs data in the real world using administrative groundwater levels data acquired from the UPGWD departmental website (UPGWD, 2020). The goal is to classify the missingness mechanism in the data and to ascertain how the deletion of missing values due to dry wells can impact estimation, inference and policy.

2.4.1 Statistical tests for missingness

We have two partially observed groundwater level variables (PREMONSOON and POSTMONSOON) and 7 constructed and fully observed climatic and land-use covariates (L0RF, L1RF, L0T2M, L1PBUILTUP, L1PDTC, L1PKHARIF, L1PRABIZAID) so we

Table 2.6: Test for MCAR, Homogeneity of Means
POSTMONSOON. Alt. hypothesis $\bar{X}_{M_i=1} \neq \bar{X}_{M_i=0}$

X	L1RF	L1PBUILTUP	L1PDTC	L1PKHARIF	L1PRABIZAID
$\bar{X}_{M_i=1}$	0.92	6.95	51.3	7.41	13.52
$\bar{X}_{M_i=0}$	0.91	7.68	52.3	7.11	13.19
p-value (α = 0.05)	≈ 0	≈ 0	≈ 0	0.0016	0.0642
df	50627	25838	23710	24134	23854
$H_0 : \bar{X}_{M_i=1} = \bar{X}_{M_i=0}$	Reject	Reject	Reject	Reject	Fail to Reject

Table 2.7: Test for MCAR, Homogeneity of Means
PREMONSOON. Alt. hypothesis $H_1 : \bar{X}_{M_i=1} \neq \bar{X}_{M_i=0}$

X	L1RF	L1PBUILTUP	L1PDTC	L1PKHARIF	L1PRABIZAID
$\bar{X}_{M_i=1}$	0.92	6.97	51.7	7.54	13.28
$\bar{X}_{M_i=0}$	0.91	7.67	52.2	7.07	13.27
p-value (α = 0.5)	≈ 0	≈ 0	0.04	≈ 0	0.93
df	49910	25839	23827	23534	24155
$H_0 : \bar{X}_{M_i=1} = \bar{X}_{M_i=0}$	Reject	Reject	Reject	Reject	Fail to Reject

must conduct 14 tests for group differences in means based on whether groundwater level observations were missing or not missing. There were significant differences in the means of 4 out of 5 covariates where observations were available versus where they were missing which rules out missingness completely at random (MCAR). Note we have fewer degrees of freedom in the case of land-use variables because the data is only available for the period 2005-2014. The results are summarized in Tables 2.4 and 2.5. We find that both the mean percentage built-up area and the mean percentage area under multiple cropping is considerably lesser where observations are missing. These findings do not in themselves provide a causal explanation but a significant difference in grouped means is suggestive evidence against MCAR whether a known causal link exists or not.

In the next series of tests, we analyze groupwise differences in means of the fully observed variables, grouping them by the *causes* of missingness in groundwater level observations. There are significant differences in grouped means for lagged monsoon rainfall, percentage builtup area and kharif area for both pre-monsoon and post-monsoon observations. For pre-monsoon, their are significant differences in means for percent-

age multiple cropped area as well. Figure 2.4 shows the plots of grouped means of the fully observed variables grouped by the causes of missingness in both PREMONSOON and POSTMONSOON observations. The relative differences between groups are generally similar for both periods of observation except for the case of lagged monsoon rainfall. We find that in pre-monsoon "closed' wells were situated in areas of relatively lower rainfall whereas for post-monsoon the dry wells were situated in areas of relatively lower rainfall as compared to wells with available observations. For both periods, choked wells correlate with higher builtup area indicating that wells located in urban settlements are more likely to register missing observations on account of becoming choked or obstructed. Kharif (rainfed) farming is generally higher in areas where wells show missingness due to unknown or other miscellaenous reasons but is lower in areas where wells are dry or closed. Multiple cropping was significantly lesser around dry wells whereas the mean of the percentage area under *Rabi* or *Zaid* (dry season) farming was higher but not significantly higher.

2.4.2 Logistic regression: identifying systemic variation in missingness

Table 2.8 lists summary statistics. Figure 2.6 shows the correlation plot for the covariates. Regression results are summarized in Table 2.9.

Lower rainfall is associated with higher log-odds of missingness irrespective of well type and the association is 10 times greater for shallow wells than that for deep wells. An increase in total Monsoon rainfall in the previous year by a depth of 1 mm spread over a neighboring area of 81 square kilometers (an increase of 81,000 cubic meters of rainwater by volume) is associated with a reduction of 0.0001 in the log-odds of missingness for DEEP wells and 0.001 for SHALLOW wells.

The stronger association of rainfall with missingness in SHALLOW wells is consistent with the hypothesis that dryness of wells drives missingness. Shallow wells are usually manually constructed and in the alluvial plains of UP, they typically do not reach depths beyond 15 m. Deep wells are tubular structures drilled into the soil either manually or using a powered drilling technology. They can reach depths from 60 m to

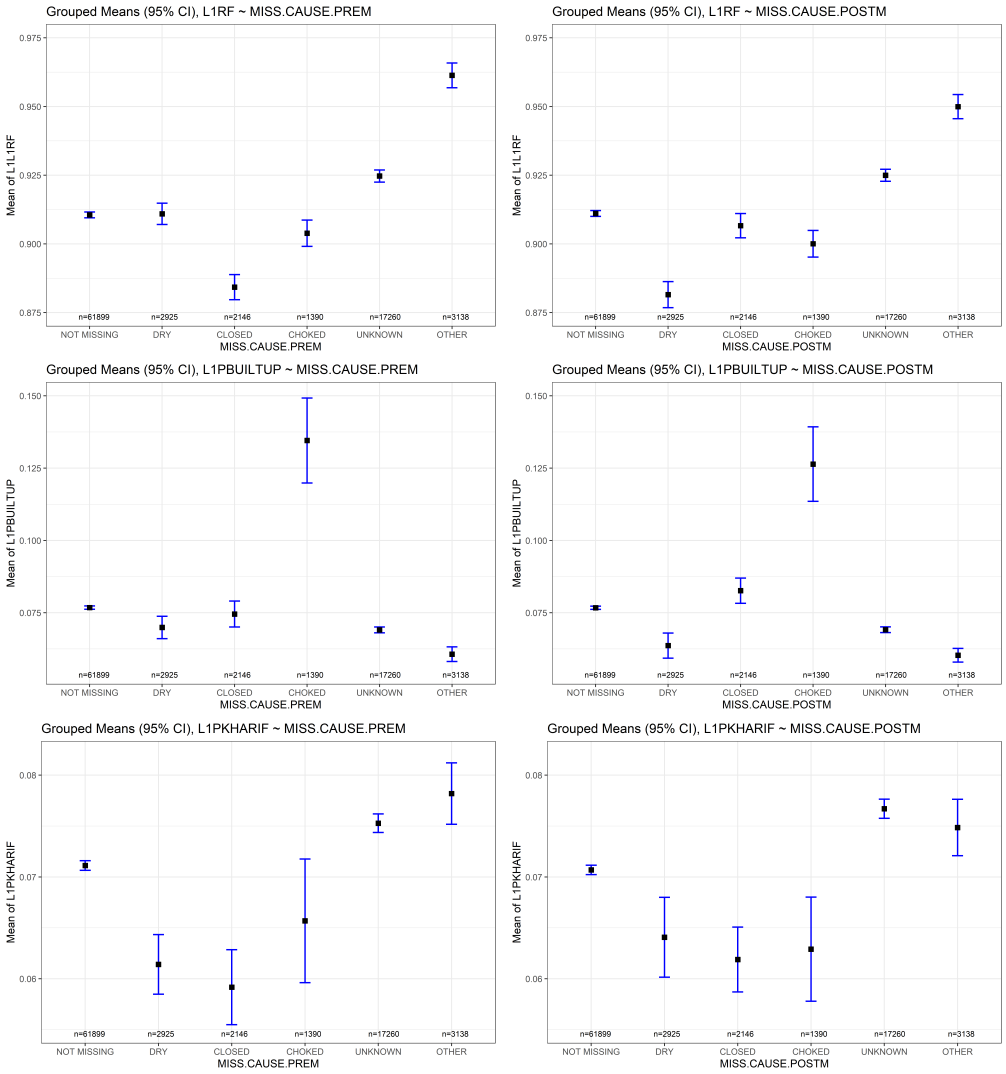


Figure 2.4: Groupwise means for fully observed variables grouped by causes of missingness for both PREMONSOON (left column) and POSTMONSOON (right column).

Variable (Units)	*Statistic*	DEEP (N = 56,486)	SHALLOW (N = 29,889)
Rainfall	*Mean*	898	927
(mm)	*SD*	262	258
	*IQR**	[721,1052]	[765,1074]
Temperature	*Mean*	297.05	297.50
(deg. Kelvin)	*SD*	1.37	1.14
	IQR	[296.10,298.03]	[296.70,298.32]
Runoff	*Mean*	2.78	2.95
(meters)	*SD*	1.19	1.11
	IQR	[1.97,3.48]	[2.15,3.60]
Multiple Crop	*Mean*	0.55	0.53
(proportion area)	*SD*	0.26	0.25
	IQR	[0.37,0.76]	[0.35,0.73]
Dry Season Crop	*Mean*	0.11	0.11
(proportion area)	*SD*	0.17	0.16
	IQR	[0.01,0.13]	[0.01,0.14]
Built-up	*Mean*	0.08	0.08
(proportion area)	*SD*	0.12	0.11
	IQR	[0.02,0.09]	[0.02,0.09]
Population	*Mean*	66	68
(1000 people)	*SD*	31	31
	IQR	[47,81]	[49,83]
Dist. to river	*Mean*	12	12
(km)	*SD*	9	9
	IQR	[6,15]	[6,16]
Dist. to highway	*Mean*	5.9	5.8
(km)	*SD*	3.8	3.8
	IQR	[3.1,7.9]	[3.0,7.8]

Table 2.8: Summary Statistics, Unit of Analysis: Well, Period: Year, 2009-2019. *[25%,75%]

	Dependent variable:		
	$ln(\frac{p(m_{i,t}^{prem}=1\|\boldsymbol{Z})}{1-p(m_{i,t}^{prem}=1\|\boldsymbol{Z})})$		
	(DEEP)	(SHALLOW)	(ALL)
Rainfall	-0.0001^{*}	-0.001^{***}	-0.001^{***}
	(0.0001)	(0.0001)	(0.0001)
Temperature	0.110^{***}	0.012	0.083^{***}
	(0.007)	(0.012)	(0.006)
Runoff	0.077^{***}	0.105^{***}	0.092^{***}
	(0.009)	(0.013)	(0.007)
Multiple Cropping	0.0003	-0.008^{***}	-0.003^{**}
	(0.001)	(0.002)	(0.001)
Dry Season Cropping	-0.002^{**}	-0.009^{***}	-0.003^{***}
	(0.001)	(0.001)	(0.001)
Built-up	-0.002^{**}	0.0001	-0.002^{**}
	(0.001)	(0.001)	(0.001)
Dist. to nearest river	-0.003^{***}	-0.008^{***}	-0.005^{***}
	(0.001)	(0.001)	(0.001)
Dist. to nearest highway	0.015^{***}	0.003	0.012^{***}
	(0.002)	(0.003)	(0.002)
Population	-0.002^{***}	-0.002^{***}	-0.002^{***}
	(0.0003)	(0.0005)	(0.0003)
Rainfall:Multiple Cropping	0.00000	0.00001^{***}	0.00000^{***}
	(0.00000)	(0.00000)	(0.00000)
Constant	-33.733^{***}	-3.527	-25.311^{***}
	(2.126)	(3.461)	(1.778)
DOF	56,486	29,889	86,375
Log Likelihood	$-33{,}928.950$	$-17{,}687.340$	$-51{,}740.790$
Akaike Inf. Crit.	67,879.890	35,396.680	103,503.600
Note:			*p<0.1; **p<0.05; ***p<0.01

Table 2.9: Logistic regression results. Separated by well type (1 DEEP,2 SHALLOW) and pooled (3). Coefficients indicate the change in log-odds of pre-Monsoon missingness with a unit change in the corresponding covariate.

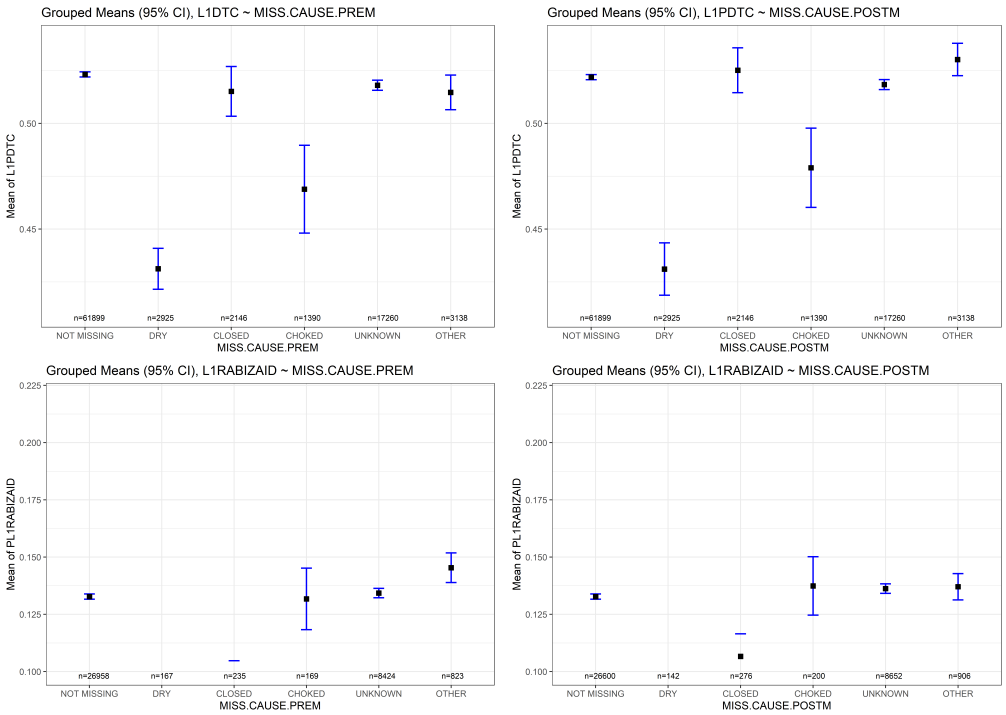


Figure 2.4: Groupwise means for fully observed variables grouped by causes of missingness for both PREMONSOON (left column) and POSTMONSOON (right column).

600 m (S. Das, 2011). Shallow wells will go dry when GWL exceeds 15 m while the first deep well will go dry when GWL crosses 60 m. We thus expect that dryness in shallow wells to be much more sensitive to rainfall deficit as compared to deep wells. Hydrological investigations using environmental tracers in parts of the Gangetic basin over eastern UP have shown that rainfall is the dominant source groundwater recharge for parts of the aquifer at depths of 40 m or less (Lapworth et al., 2021). The higher sensitivity of log-odds of missingness to variation in rainfall for shallow wells accords with these hydrological findings.

The percentage area under multiple cropping within 1 km radius of an observation is not associated with the log-odds of missingness of deep wells but it is associated with a reduced missingness in shallow wells. A 1% increase in multiple cropped area around a shallow well is associated with 0.008 less missingness. Conversely, an increase in log-odds of missingness by 0.008 is associated with a 1% lower multiple cropped area. The converse interpretation accords with the hypothesis that missingness occurs due to dry wells. In areas with a greater number of dry wells, we expect lesser multiple cropping since it requires higher groundwater input. The same applies to dry season cropping

which is mostly irrigated. Increase in log-odds of missingness by 0.002 for deep wells and 0.009 for shallow wells are associated with a 1% reduction in area under dry season cropping. Increase in built-up area by 1% is associated with a reduction in missingness in deep wells and not associated strongly for shallow wells.

The interaction of rainfall and multiple cropping is significant for shallow wells but not deep wells. In shallow wells, an increase in rainfall by 1 mm alone is associated with a reduction in log-odds of missingness by 0.001 but if additionally, there is an increase in multiple cropping by 1%, then the negative association with log-odds is reduced (or weakened) by 0.00001. This suggests that for shallow wells only, while rainfall alone leads to a reduction in missingness, the effect is weaker if multiple cropping is relatively higher.

Higher temperature is associated with higher log-odds of missingness overall and for deep wells but not for shallow wells. In general, hotter climate should be associated with more dryness therefore with higher missingness according to our hypothesis. The fact that we find the opposite is true for shallow wells is partially explained by the fact that shallow wells tend to be concentrated in colder regions in the northern and eastern parts while deep wells in the hotter southern and western regions of UP (see Figure 4.2.2).

Higher groundwater runoff is consistently associated with higher log-odds of missingness across well types and overall. Higher runoff reduces the amount of groundwater recharge and we expect it to contribute to depletion in GWL and thus to greater log-odds of dryness.

We find the log-odds of missingness to be lower nearer to major rivers and distributaries. The Ganges river, which is the major river of the basin, is classified as "gaining" river[23] that receives water from adjoining groundwater aquifers through "baseflow" particularly during the non-Monsoon, dry months (A. Mukherjee et al., 2018). Greater missingness due to dryness near rivers could be attributed to such loss of water to rivers and/or to heavier population density and economic activity on the extremely fertile Gangetic floodplains.

[23]Rivers that lose water to neighboring aquifers are called "losing" rivers.

The log-odds of missingness are lower near major highways for deep wells but no significant effect is found with respect to highway proximity for shallow wells. The log-odds are, on average, less in more populous areas compared to sparsely populated ones. These effects could be attributed to more regular monitoring near highways and towns as compared to the interiors rural areas of the state and may be unrelated to dryness.

2.4.3 Estimated probabilities of missingness

Using estimated parameters from our logistic regression, we predicted the probabilities of missingness for each well observation. Since we have chosen covariates that specifically drive the process of drying up of wells, we expect that the predicted probabilities of missingness will be greater for those missing well observations that were labeled "DRY" as compared to those that were not missing. The grouped means of fitted probabilities by missingness label are shown in figure 2.5. We do find that the mean probability of missingness is significantly greater for wells labeled "DRY" as compared to wells with no missingness both for shallow and deep wells.

For wells where the missingness was due to unknown causes, we find that the probability of missingness is similar to wells labeled "DRY" in the case of shallow wells. For the case of deep wells, the mean predicted probability for observed values and for missing values due to unknown causes are similar. This suggests that a significant proportion of missing values where the cause was unknown occurred due to dry wells for shallow wells but not for deep wells.

Well observations that were missing due to "OTHER" causes also had a relatively higher mean predicted probability implying that some of those may also have been dry. Note that the causes of missingness are generally not mutually exclusive. A well marked "DAMAGED" (which would appear as part of the "OTHER" category) may also have been dry. The only mutual exclusions exists between observed and dry wells. A well where GWL was observed could not possibly have been dry and conversely, a dry well could not possibly be observed.

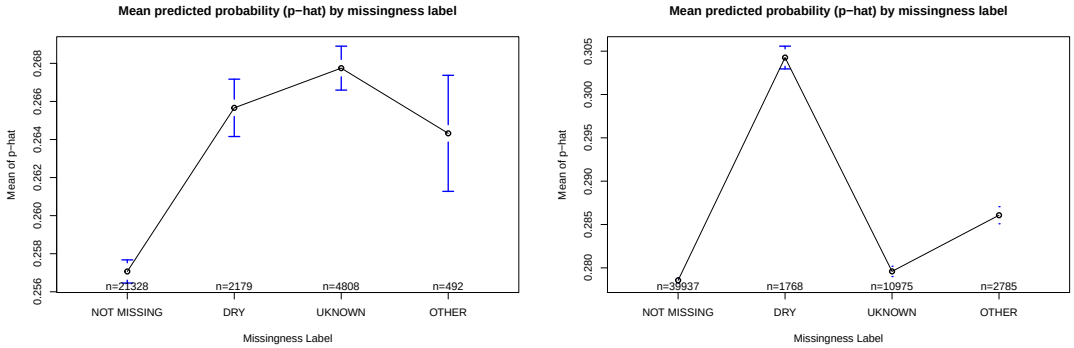


Figure 2.5: Mean estimated probabilities by missingness label (left: SHALLOW wells, right: DEEP wells). The labels indicate the cause of missingness as reported by the agency.

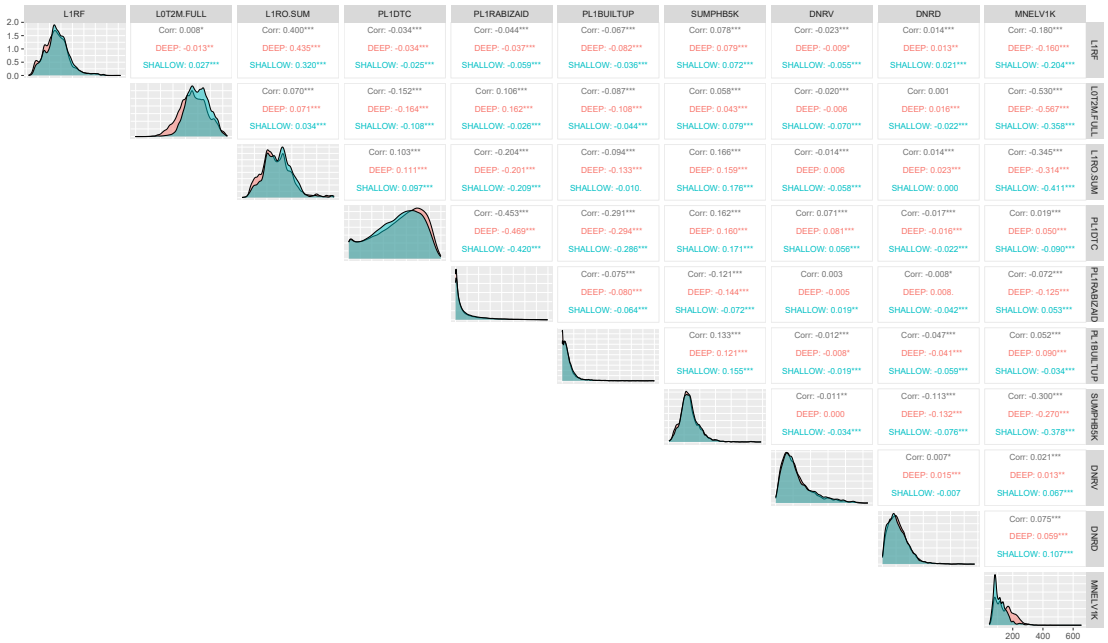


Figure 2.6: Diagonal: density plots of covariates , Upper: correlations between covariates. Grouped by well type (DEEP, SHALLOW). L1RF (Rainfall), L0T2M (Temp.), L1RO (Runoff), PL1DTC (Mult. Crop), PL1RABIZAID (Dry Season Crop), PL1UILTUP (Built-up), SUMPHB5K (Population), DNRV (Dist. to rivers), DNRD (Dist. to highways), MNELV1K (Elevation).

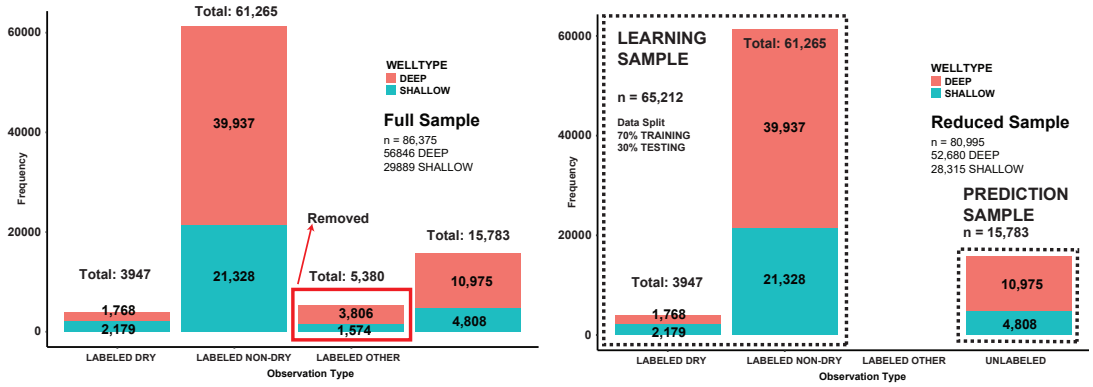
2.4.4 Predictive Model for Unlabeled Missing Data

We are encouraged by the fact that the results from the logistic regression model lends weight to our hypothesis that dry wells are a major underlying cause of missingness as indicated by signs and magnitudes of estimated coefficients discussed in the previous section. It is also encouraging that the mean estimated probabilities of missingness grouped by missingness labels are higher for wells labeled DRY compared to wells that were observed for both deep and shallow wells (see Figure 2.5). Note also that Figure 2.5 shows that the mean probability of missingness for unlabeled missing data is close to that for DRY labeled missing data. This suggests that dryness of wells is also a major cause for unlabeled missing data. This is not so for deep wells where the mean probability of missingness is much lower for unlabeled missing data than for DRY labeled data. This makes sense since we expect shallow wells to dry up more easily and more frequently whereas missingness deep wells dry up less easily and less frequently and it is likely that missingness in deep wells is more due to other random causes. We use the model specification given by equation 2.14 to build a predictive logistic regression model for unlabeled missing data.

The steps to build the predictive logistic regression model are described in section 4.4.2. The model specification remains the same but now we work only with labeled data. Well observations where the GWL is observed (not missing) are implicitly labeled "NON-DRY" since observed wells cannot be dry. Well observations where GWL is missing due to dry are labeled DRY. We remove all other observations from the sample (step 1 in listing 4.4.2). We then proceed according to steps 2-7 for 100 Monte Carlo iterations.

2.4.5 Bias Estimation

The predictive model provides a prediction which is either "DRY" or "NON-DRY" for each missing observation. Using these predicted labels, we augment the GWL data by substituting the maximum well depth $\bar{g}$ in place of the missing value. The data set thus augmented is still not a complete representation of the real world for three reasons. First, due to confusion in the predictions. Second, due to uncertainty in selection of



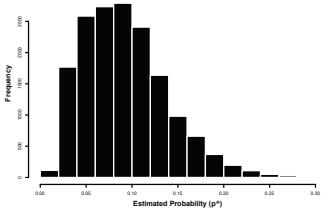
(a) Full Sample

(b) Reduced Sample

	Dependent variable:
	log-odds of missingness
Rainfall	−0.002***
	(0.0003)
Temperature	0.311***
	(0.026)
Runoff	0.308***
	(0.026)
Multiple Crop	−0.019***
	(0.004)
Dry Season Crop	−0.006**
	(0.002)
Built-up	0.004
	(0.003)
Dist. to nearest river	−0.012***
	(0.003)
Dist. to nearest highway	0.018***
	(0.007)
Population	0.001
	(0.001)
Rainfall:Multiple Crop	0.00002***
	(0.00000)
Constant	−94.455***
	(7.898)
Observations	16,455
Log Likelihood	−4,876.302
Akaike Inf. Crit.	9,774.604
Note:	*p<0.1; **p<0.05; ***p<0.01

(c) Training Results (SHALLOW wells)

	Statistic $\hat{p}$	Value
1	Min	0.005444
2	1st Qu.	0.057346
3	Median	0.087232
4	Mean	0.092677
5	3rd Qu.	0.119249
6	Max	0.314887



(d) Fitted Probabilities

Bias Estimation	
#Labeled DRY: nld	2179
#Labeled NON-DRY: nlnd	21328
#Total Labeled:	23,507
#Predicted DRY: npd	1,545
#Predicted NON-DRY: npnd	3,263
#Total Predicted:	4,808

$\hat{d}$ = nld + npd = 3,724
$\hat{d}_{lower}$ = nld + rtp*npd = 2,380
$\hat{d}_{upper}$ = 3,822
max depth = 15 meters

$\hat{\mu}_{trunc}$ = 8.21 m
$\hat{\mu}_{lb}$ = 9.23 m
$\hat{\mu}_{lb,\ lower}$ = 8.9 m
$\hat{\mu}_{lb,\ upper}$ = 9.27 m

Bias interval [0.69, 1.06] m

(e) Bias Estimation

Predicted Label →

True Label ↓	DRY	NON-DRY
DRY	420	234
NON-DRY	2,720	3,678

True Positive Rate: r_{tp} = **0.13**
False Negative Rate: r_{fn} = **0.06**

(e) Confusion Matrix

Monte Carlo Simulation for Bias Estimation (case of SHALLOW wells)

STEP 1: Remove Random Missingness
Remove well observations that were missing and labeled due to "other" causes (other than dry wells). We assume that this missingness is random and the removal does not affect inference. The full sample (a) is thus reduced (b).

STEP 2: Data Split
Separate data into the "learning" sample which contains the labeled data. The labels are DRY (1) and NON-DRY (0). The remaining part is unlabeled data for which we predict labels. Split labeled data into training (70%) and testing (30%) split.

STEP 3: Training logistic regression model
Use training split to run the logistic regression model. Results are summarized in (c). Use the estimated coefficients to predict a probability of missingness for each observation. Figure (d) shows the summary stats/histogram for estimated probabilities.

STEP 4: Confusion Matrix
We make predictions using the first quartile, mean and third quartile of the estimated probabilities (see table in Figure (d), top) as the threshold of prediction. We use the predicted labels to build a confusion matrix (e). This matrix is for the mean p*(red).

STEP 5: Bias Estimation
Using the predicted labels, for each well that is either labeled or predicted DRY, we substitute the maximum depth (15 m for SHALLOW wells)

We then compute the truncated and predicted spatial means. We also compute upper and lower limits for the predicted mean.

Finally, we compute an interval for the minimum bias due to removal of DRY wells.

STEP 6: Monte Carlo Iteration
Repeat Steps 2-5 for 100 iterations and report the averages and confidence intervals. We report these results later.

Figure 2.7: Monte Carlo Simulation Method for Unconditional Spatial Mean Bias Estimation.

the probability threshold of the predictive model and finally, due to the fact that the "imputed" value $\bar{g}$ is only a lower bound for the actual GWL at the well. The GWL at a dry well will always be greater than $\bar{g}$. We argue in the following sections that the augmented data still possesses a utility in assessing the impact of non-random missingness on statistical inference. We utilize the augmented data to obtain an approximation of the bias in point estimates of the unconditional spatial mean and conditional mean GWL.

Unconditional Spatial Mean

In the case of the spatial mean, our augmented data provides an indication of the minimum bias that will occur due to deletion of missing values in dry wells. The minimum bias thus obtained will be subject to confusion due to type I and type II errors in the predictive logistic regression model. A type I error (a false positive) occurs if a well that was actually not dry is predicted as being dry. A false positive prediction will result in over-estimation of the truncation bias since. A type II error (a false negative) occurs when a well that will actually dry is predicted as not dry. This will result in under-estimation of truncation bias. We place bounds on the truncation bias estimates by using the confusion matrix of the logistic regression as explained in section 4.4.2. We report the predicted truncation biases for three different values of p^* to assess the impact of the choice of p^* on results.

Figure 2.7 describes the entire procedure for the Monte Carlo simulation visually for the case of SHALLOW wells (the case of DEEP wells proceeds in the same way). The full sample contains 86,375 observations (56,486 DEEP and 29,889 SHALLOW). The removal of missing observations that occurred due to causes other than dry wells (as indicated by their labels) leaves a reduced sample of 64,842 observations (38,587 DEEP and 26,255 SHALLOW) (see Figure 2.7, panels (a) and (b)). In the reduced sample, we had 3,947 missing observations that were confirmed dry wells as per their labels and 61,265 observed GWL values that we label "NON-DRY". The remaining 15,783 observations were missing without labels and our goal from this prediction exercise is to see how many of these were due to dry wells.

We split the labeled data (65,212) using a 70-30 split with 70% in the training set

and 30% in the testing set (Vrigazova, 2021). We train the logistic regression model using our labeled data with a value of 0 if well is NON-DRY and 1 if DRY. Figure 2.7 (panel (c)) summarizes the results of training. Estimated coefficients and signs resemble the main regression. Using these estimates, we estimate the probability of well dryness for each missing observation $\hat{p}_i$. The summary stats are given in Figure 2.7 (panel (d), top). Estimated probabilities range between approximately 0.005 and 0.315. This is relatively lower compared to the typical logistic regression where the mean estimated probability is around 0.5. The reason for low values is unbalanced category distribution. Only 6% of the training data correspond to the value 1 (dry wells). A low proportion of 1s (or a high proportion of 0s) is not necessarily a source of bias in logistic regression. Salas-Eljatib (2018) discuss the impact of unbalanced data on logistic regression estimates. The authors show that nearly unbiased estimates can be achieved with a category distribution in which 90% of the dependent variable values are 0s (a case similar to ours) but the variance of the estimators increases with extreme category distributions (Salas-Eljatib et al., 2018).

We assess the bias in spatial mean estimates using three different probability thresholds p^*. We use the first quartile (Q1), the mean and the third quartile (Q3) of the distribution of fitted probabilities (shown in blue and red font in Figre 2.7, panel (d)). The results in the figure are for the mean probability (p^* = 0.092677) which we approximate as 0.09. We predict a well with a missing observation as DRY if the fitted probability for that well $p_i > p^*$ and NON-DRY otherwise. This results in the confusion matrix shown in Figure 2.7 (panel (e)). The precision r_{tp} is 0.13 and the false negative prediction rate is 0.06. These results indicate fairly high confusion in the prediction and we attribute this to the increase in variance of estimators in the predictive logistic regression due to unbalanced data. King and Zeng (2001) demonstrate methods to achieve precise and accurate inference for "rare events data" where a particular category (0 or 1) predominates (King & Zeng, 2001). The integration of such methods is the subject of future work. At present, we rely on the fact that the significance and accuracy of estimators is robust to randomized data splits across 100 Monte Carlo iterations.

Finally, we obtain (for the case shown in Figure 2.7, see panel (e)) 1,545 predicted DRY wells along with 2,179 labeled DRY wells and 3,263 predicted NON-DRY wells

along with 21,328 labeled NON-DRY wells. We discard the missing data for wells that were predicted NON-DRY. For each DRY well (labeled or predicted), we substitute a maximum depth of 15 m for SHALLOW wells in place of the missing GWL observation. We retain labeled NON-DRY wells as they are since we observe the values of GWL for those. The labeled NON-DRY wells where GWL is observed constitute our truncated data containing 21,328 observations. Our augmented data containing labeled and predicted DRY wells where observations were originally missing contains 3,724 observations. The augmented dataset thus contains 21,328 + 3,724 = 25,052 observations.

The truncated sample mean ($\hat{\mu}_{trunc}$) was found to be 8.21 m while the lower-bound sample mean ($\hat{\mu}_{lb}$) computed using the augmented data was 8.9 m. Accounting for confusion, the upper bound of the lower-bound mean ($\hat{\mu}_{lb,lower}$) was 8.9 m while the upper bound ($\hat{\mu}_{lb,upper}$) was 9.27 m. The estimated bias in the spatial mean thus lay in the interval [0.69 m, 1.06 m]. The results averaged over 100 iterations are shown in Figure 2.10 for SHALLOW wells (left panel) and DEEP wells (right panel). The plots show the expected values of the truncated (solid black line) and the lower-bound (solid blue line) sample means along with the upper and lower limits (purple dotted lines) of the lower-bound sample mean (on the y-axis) as a function of the probability threshold p^* (on the x-axis). The figure also shows the interval of estimation bias for each value of p^*.

As can be seen in the figure, regardless of p^*, the envelop of the lower-bound mean (pink shaded region) is distinct from the truncated mean. As p^* increases, we predict fewer DRY wells and as expected, the bias decreases and the lower bound sample mean tends towards the truncated mean. The bias for DEEP wells is greater than that for SHALLOW wells. The bias for DEEP wells is unusually high for $p^* = 0.01$ which we attribute to the high rate of false positives. The results for DEEP wells are generally subject to greater uncertainty due to extreme imbalance in the category distribution with dry wells (category 1) being only 4.5% of the total number of wells.

Conditional Mean

Table 2.12 list summary statistics for the dependent variable (Pre-Monsoon GWL) and rainfall covariates that appear in regression specified by equation 2.25. Note that the maximum value of GWL (1915 m) is absurdly high. We consider it an outlier since it does not correspond to any relatable physical reality. The boxplot in Figure 2.8 (top-left panel) shows two outliers which we remove. The plot after removal is shown in Figure 2.8 (top-right panel).

Since GWL is a positive quantity (the minimum value 0 mbgl corresponds to the groundwater being just under the ground surface), a normal distribution assumption for the data is not appropriate. We instead fit a gamma distribution to the GWL data using maximum likelihood estimation (MLE) after removing outliers and missing values. Figure 2.8 show (clockwise from top-left panel) the empirical histogram and fitted theoretical probability density curve, the Q-Q plot, the P-P plot and the empirical and theoretical CDFs. For comparison, Figure 2.9 plots theoretical densities of the best fits of the normal and exponential families of distributions and the best fit of the gamma distribution against the density histogram of the GWLs data. The gamma distribution appears to be the best fit although the Q-Q plot (Figure 2.8 middle panel) deviates from the theoretical estimate between the values of 20 mbgl and 50 mbgl. Table 2.11 summarizes the results of the fitting procedure and the estimated shape and rate parameters (shape = 2.59, rate = 0.014) of the fitted gamma distribution. We estimate the parameters $\vec{\eta}$ from equation 2.25 using MLE with a gamma distribution assumption based on our best-fit result.

The Monte Carlo simulation procedure follows a similar procedure to that followed for the unconditional mean estimation (steps as described in section 4.4.2). Once DRY well predictions are made on the unlabeled data, for each well that is predicted DRY, we substitute $g_i = \bar{g}$, i.e we replace the missing GWL observation with the maximum depth of the well depending on type (15 m for SHALLOW and 60 m for DEEP). We obtain estimates for $\vec{\eta}$ by first truncating all missing observations and using only the observed data ($\hat{\eta}_{trunc}$) and then by making substitutions for wells predicted DRY ($\hat{\eta}_{pred}$). We then compute the bias as the difference between the two estimates. We compute expected values by Monte Carlo iteration.

To obtain $\hat{\eta}_{pred}$, we make the substitution $g_i = \bar{g}$ although this is just a lower bound. In reality, the GWL for DRY wells may be further below (i.e greater in magnitude) than $\bar{g}$ thus the estimate $\hat{\eta}_{pred}$ is obtained under an assumption of a "best-case" GWL scenario where DRY wells are *just barely* dry. As before, we obtain the estimates of $\hat{\eta}_{pred}$ for three different values of p^* corresponding to the first quartile, mean and third quartile of $\hat{p}$. We repeat the estimation procedure for 100 Monte Carlo iterations for each well type (SHALLOW and DEEP) and then compute expected values and standard deviations by averaging across the obtained estimates. Table 2.10 summarizes the simulation results.

Estimates obtained with the truncated data (Table 2.10, column 2) are sensible. The coefficients for normal and surplus rainfall have negative signs indicating that an increase in rainfall is associated with decrease in GWL magnitude suggesting the occurrence of greater aquifer recharge that brings water level closer to the surface (and thus causes depth to decrease). The coefficient for rainfall deficit is positive suggesting that rainfall deficits, on average, cause GWL to increase (or deepen). The same signs are recovered when data are augmented through DRY well prediction. We thus do not find that truncation of missing data due to dry wells causes a reversal of sign for any coefficient in this model.

We do find that, as expected, the truncated regression increases uncertainty suggesting information loss. Standard errors reduce consistently after data augmentation for all values of p^* and for all coefficients. Standard errors start rising again as the probability threshold rises. This is due to the fact that when the threshold rises, we predict fewer and fewer dry wells and the regression approaches the truncated regression. The coefficient of surplus rainfall ($\hat{\eta}_2$) is not significant (i.e its confidence level is less than 90%) in the truncated regression but it becomes significant at the 99% confidence level after data augmentation (at p^* = 0.06). This is suggestive evidence that truncation of missing values due to dry wells may obscure the relationship between rainfall and GWL. The coefficient for rainfall deficit ($\hat{\eta}_3$) is consistently significant but smaller in magnitude in augemented regression as compared to the truncated regression.

Overall, we have found that for our model, the truncation of missing values due to dry wells did not alter the inferred relationship between normal (or average) rainfall and GWL but it obfuscated the one between rainfall variation and GWL. The impact

Coeff.	Truncated	Predicted			% Truncation Bias		
		p^*=0.06	p^*=0.09	p^*=0.12	0.06	0.09	0.12
$\hat{\eta}_1$	-1.1×10^{-3}***	-9.8×10^{-4}***	-9.5×10^{-4}***	-9.7×10^{-4}***	11	15	13
	(2.9×10^{-5})	(2.42 ×10^{-5})	(2.57 ×10^{-5})	(2.63 ×10^{-5})	-20	-14	-11
$\hat{\eta}_2$	-3.9×10^{-5}	-1.5×10^{-4}**	-1.0×10^{-4}	-1.2×10^{-4}	-74	-62	-66
	(3.9×10^{-5})	(3.16 ×10^{-5})	(3.40 ×10^{-5})	(3.52 ×10^{-5})	-24	-15	-11
$\hat{\eta}_3$	1.3×10^{-4}**	9.2×10^{-5}*	1.9×10^{-4}***	2.0×10^{-4}***	-41	30	35
	(4.8×10^{-5})	(3.7×10^{-5})	(4.0×10^{-5})	(4.2×10^{-5})	-30	-20	-14
.$p < 0.1$, *$p < 0.05$, **$p < 0.01$, ***$p < 0.001$							

Table 2.10: Maximum Likelihood Estimates of Regression Coefficients for the Conditional Mean GWL. $\hat{\eta}_1$ is the coefficient of normal rainfall at well i, $\hat{\eta}_2$ for annual surplus rainfall at well i and $\hat{\eta}_3$ for annual deficit in rainfall at well i. Prediction is made using three different values (0.06, 0.09, 0.12) of threshold probability p^* corresponding to the first quartile, mean and third quartile cutoffs of the estimated probabilities.

Parameter	Estimate	Std. Error
shape	2.59	0.014
rate	0.29	0.002
n: 61,263		
log-likelihood: -183733.4 **AIC**: 367470.8 **BIC**: 367488.9		
Correlation matrix:		
	shape	rate
shape	1.00	0.91
rate	0.91	1.00

Table 2.11: Estimated shape and rate parameters and goodness of fit for Gamma distribution using MLE.

of surplus rainfall and droughts (or deficits) on GWL is an important policy concern. Climate change is causing rainfall variations to move towards extremes (Arora & Ali, 2022). The increase in the intensity and frequency of occurrence of flood and drought stands to worsen the conundrum posed by non-ignorable missing data.

Statistic	N	Mean	St. Dev.	Min	Max
Pre-Monsoon GWL (mbgl)	61,265	9.034	10.103	0.030	1,915.000
Normal Rainfall (mm)	80,995	905.084	146.934	599.087	1,959.744
Annual Surplus (mm)	80,995	86.389	138.225	0.000	985.606
Annual Deficit (mm)	80,995	84.452	118.580	0.000	592.094

Table 2.12: Summary statistics, GWL and Rainfall covariates.

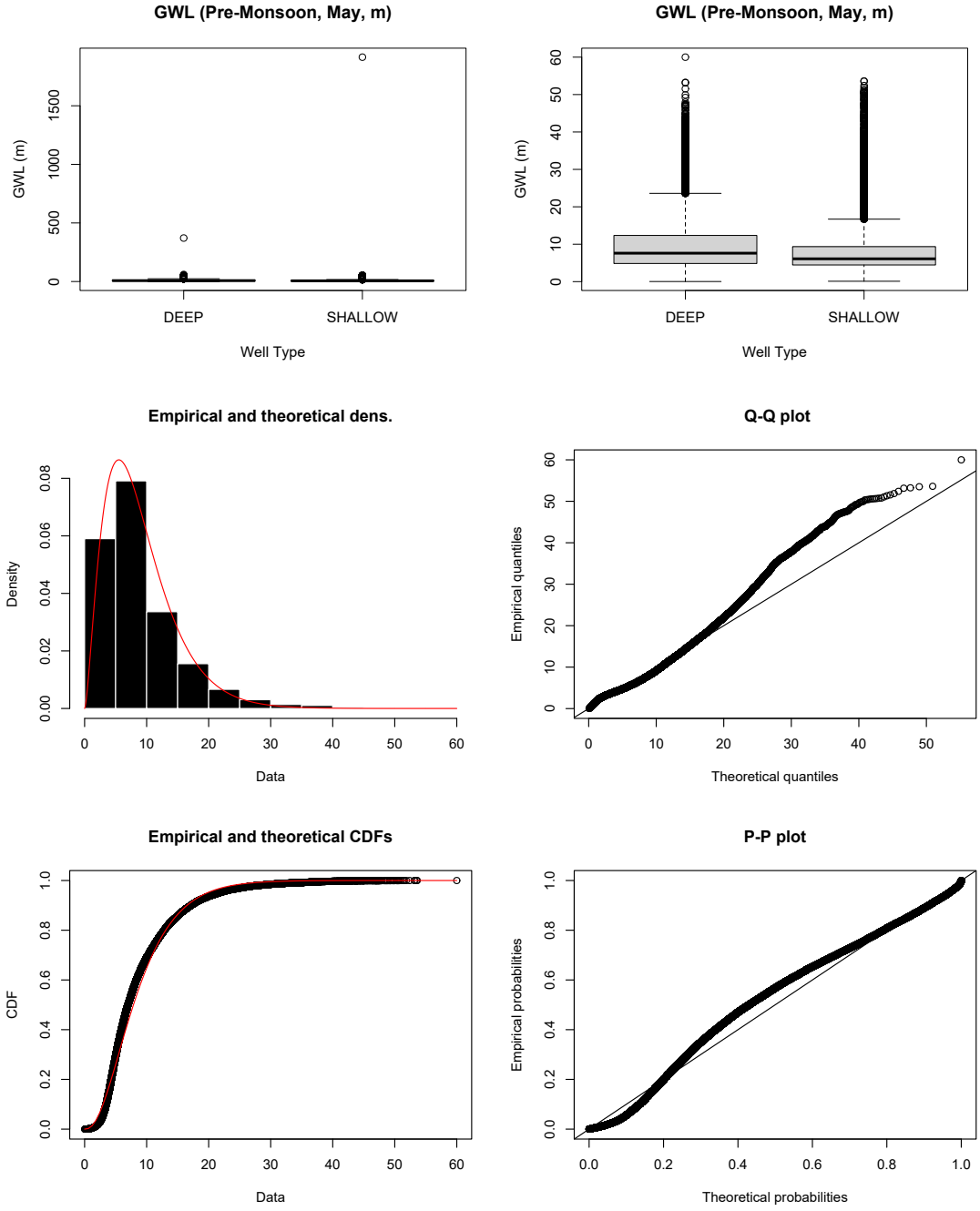


Figure 2.8: Top: Boxplots of Pre-Monsoon GWL with outliers (left) and without outliers (right). Bottom: Fitting a gamma distribution to the GWL data after removing outliers and missing values.

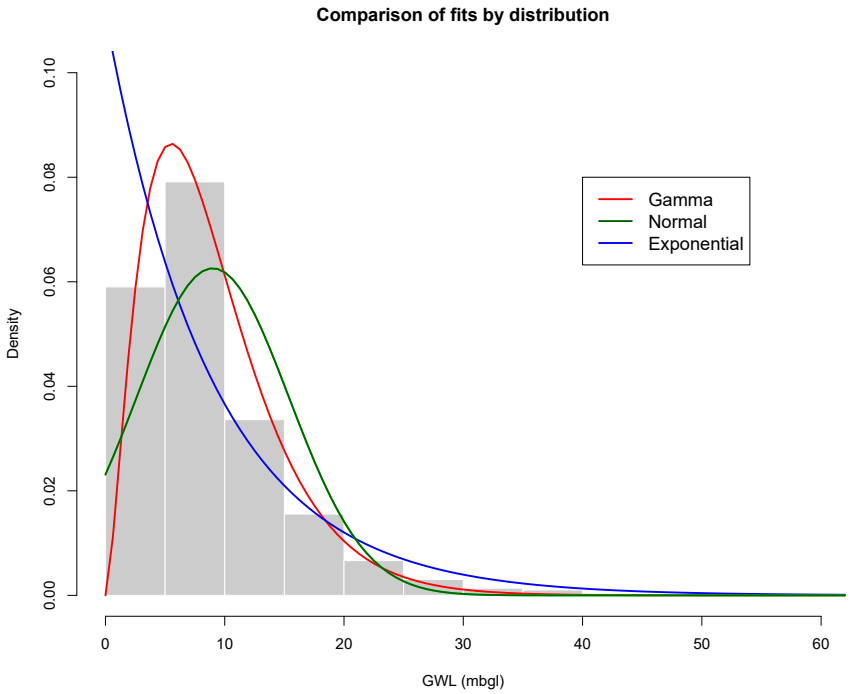


Figure 2.9: Comparison of different distributions as fitted to the GWLs data. The red line indicating the theoretical PDF of the Gamma distribution appears to be the best fit.

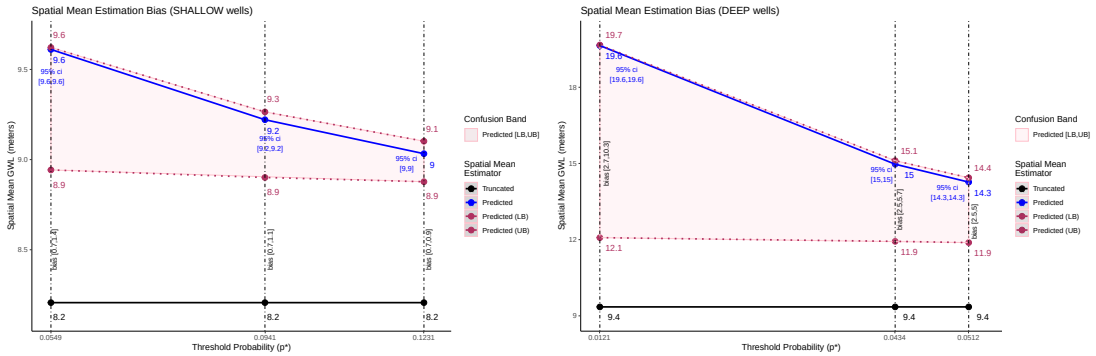


Figure 2.10: Unconditional Spatial Mean Estimation Bias (Left: DEEP wells, Right: SHALLOW wells)

2.5 Conclusion

Dry wells are qualitative indicators of water scarcity (Varghese et al., 2013; USGS, 2018b). They carry important information about groundwater availability in a region (Hora et al., 2019). We have shown that their omission from quantitative models is a methodological limitation that incurs a loss of information. Further, we give evidence that in the Gangetic Basin of Northern India, drought and agricultural land-use intensification exacerbate the occurence of missing data.

Using Rubin's missing data theory, we have shown with evidence from Uttar Pradesh that the process that governs dryness of wells is a non-random (MNAR) missingness mechanism. GWL data collected through monitoring wells will always have non-random missingness in the presence of dry wells. The naive deletion of missing values is frequently reported in the literature and the issue of missing groundwater levels data is not sufficiently discussed in groundwater management and policy documents. We have demonstrated the biases that will occur in the estimation of spatial unconditional means and regression coefficients when missing values due to dry wells are ignored. We have also discussed the potential impact on groundwater resource management and policy in the Indian context. While dry wells do not furnish numerical observations for groundwater levels, the knowledge of dryness can be used to quantify the extent of mis-estimation. We thus recommend that dry wells be explicitly recorded and reported along with groundwater data publications by monitoring agencies.

This study has several limitations and simplifying assumptions which present pathways to future work. While we were working with data set where the causes of a certain fraction of missing values were stated, this is not the case in general for groundwater monitoring data (Perrone & Jasechko, 2017a) and prediction of underlying missingness mechanisms is necessary for a generally applicable approach. We have restricted ourselves to univariate missingness and commented on how bivariate missingness patterns are important while computing blockwise SGEs but in general more than two variables may have missing values thus multivariate missingness mechanisms in groundwater monitoring data are an important area of future research. We studied a special case in which all covariates were fully observed and only outcomes had missingness and this

may not be the case in general. We have identified but not studied the mechanisms behind many other causes of missingness reported in the groundwater monitoring data from Uttar Pradesh that may hold important clues about the processes that cause missing values to appear.

2.6 Discussion

The attention and emphasis on groundwater resource management in India has steadily increased since economic liberalization in the early 1990s (CGWB, 2019; Ministry of Jal Shakti, 2020; UPGWD, 2019; Shah, 2010). Groundwater governance agencies notify regulation such as the restrictions on the number of groundwater abstraction and rainwater harvesting structures in an area, or the refinance of loans for tubewells based on the SGE (UPGWD, 2019; CGWB, 2019a; IRMED, n.d.). The SGE is computed within sub-district administrative units called “blocks” using administrative groundwater levels data collected at observation wells by the CGWB along with state level agencies. The SGE is defined as the ratio of the "current annual gross groundwater draft for all uses" (AGD) in a block to the "annual extractable groundwater resources available" (AGA) in that block expressed as a percentage (CGWB, 2019a). The SGE in block i is given by equation 2.28.

$$SGE_i = \frac{Current\ annual\ gross\ groundwater\ draft\ for\ all\ uses}{Annual\ extractable\ groundwater\ resources\ available} \times 100 \quad (2.28)$$

The term in the numerator - the annual groundwater draft (AGD), is computed using irrigation data, power consumption data and from other sources like standard crop water requirement data. The term in the denominator which we abbreviate as "AGA" for "annual groundwater availability" is computed using the "water-table fluctuation method" alternatively known as the "groundwater balance method." If a block i contains n observations wells $w_1, ..., w_n$ and they each yield observations $y_{1,prem}, ..., y_{n,prem}$ in pre-monsoon and $y_{1,postm}, ..., y_{n,postm}$ in post-monsoon. We then compute the average seasonal water level fluctuation in block i as:

$$\Delta\bar{g}_i = \frac{\sum_{i=1}^{n}(g_{i,prem} - g_{i,postm})}{n} \tag{2.29}$$

The change in storage of the aquifer below the assessment block is computed by averaging groundwater fluctuations across the block according to equation 2.30. Individual well-level variations in groundwater levels in an assessment year thus gets aggregated to the block level (CGWB, 2017).

$$\Delta S_i = \Delta\bar{g}_i \times S_i \times A_i \tag{2.30}$$

where ΔS_i is the total change in water storage by volume in the aquifer column beneath the assessment block, $\Delta\bar{g}_i$ is the average rise or fluctuation in water levels after the monsoon rainfall, S_i is the specific yield of the aquifer medium and A_i is the area of the assessment block. The total change in storage reflects all of the inputs and outputs to the aquifer in a monitoring season.

It is important to discuss the impact of missing observations on the SGE because it has direct policy implications. The CGWB classifies each block into a category based on the computed value of the SGE in that block. Table 2.14 lists the criteria of classification and table 2.15 lists the number of blocks in each category in the year 2017 (CGWB, 2019). Figure 2.11 shows the spatial distribution of categories in the state of Uttar Pradesh. The western and south-western parts of the state have a higher incidence of blocks in critical and over-exploited categories.

To understand how missingness impacts the SGE, we must consider bivariate missingness since both the pre-monsoon and the post-monsoon are involved (see equation 2.29). There are four possible patterns of missingness that are listed in Table 2.13 and numbered from 0 to 3. Let's assume that all missing observations occur due to wells being dry. Pattern 0 corresponds to no missingness since both the pre-monsoon and post-monsoon values are observed. If the pre-monsoon value is missing and the post-monsoon value is observed (pattern 1), it implies that the well was dry in the summer but became functional after the monsoon and thus it experienced a net positive fluctuation. If in a block, pattern 1 missingness predominates then the deletion of cases with miss-

ing values will cause an underestimation of seasonal fluctuation and thus the amount of recharge estimated for that particular block will be less than the actual recharge. The denominator in equation 2.28 will be smaller than the true value and the apparent SGE for the block will be greater than the true SGE. The block will appear to be at a more advanced stage of groundwater extraction than it really is. Removing pattern 2 missingness will have the opposite effect. If for a well, the pre-monsoon value was observed and the post-monsoon value was missing then it means that the well was functional in the summer but became dry after the monsoon. We can be sure in this case that it experienced a net negative seasonal fluctuation. Removing these wells will cause the denominator term in equation 2.28 to be greater than its true value and we will assign a value to the SGE that does not account for a certain amount of groundwater development and the block will appear underdeveloped. For the case of pattern 4 missingness where both the pre-monsoon, the true seasonal fluctuation may either be negative or positive. The well was dry in both seasons and it is possible that this missingness may be ignorable unless a systematic relationship exists between perennial dryness of wells and the seasonal fluctuation.

The CGWB uses decadal trends in groundwater level to validate their computation of the SGE for a given block. If a block is categorized as over-exploited based on the SGE but does not show a significant declining trend in water level in both pre-monsoon and post-monsoon seasons over a decade then it is marked for reassessment. Similarly, if a block is deemed safe according to the SGE but shows declining decadal trends in groundwater level, it is marked for reassessment. While this validation procedure may capture certain anomalies in SGE computation, the significance of long terms trends has also been shown to be sensitive to listwise deletion of missing values (Hora et al., 2019).

Pattern	Pre-monsoon	Post-monsoon	Fluctuation	Impact on SGE
0	Observed	Observed	+ve / -ve	-
1	Missing	Observed	+ve $(g_{i,prem} > g_{i,postm})$	Apparent overdevelopment
2	Observed	Missing	-ve $(g_{i,prem} < g_{i,postm})$	Apparent underdevelopment
3	Missing	Missing	+ve / -ve	Indeterminate

Table 2.13: Bivariate Missingness Patterns for Pre-Monsoon and Post-Monsoon Variables and the Impact on SGE by Deletion

Stage of Groundwater Extraction (SGE)	Category
$\leq 70\%$	Safe
>70% and $\leq 90\%$	Semi-Critical
>90% and $\leq 100\%$	Critical
>100%	Over Exploited

SGE	Trend (10 yrs)	Remarks
$\leq 70\%$	Decline in pre-monsoon and post-monsoon	Needs reassessment
>100%	No significant decline in pre-monsoon and post-monsoon	Needs reassessment

Table 2.14: Criteria for categorization of assessment units based on SGE. The national compilation of the dynamic groundwater resources of India based on the joint assessment of 2017 provides a criteria to classify blocks into categories based on the computed SGE (top) and also a stipulation for validating the SGE computation with long term water level trends (bottom). (CGWB, 2019)

Category	Number of AU's
Over-Exploited	91
Critical	48
Semi-Critical	151
Safe	540
Total	830

Table 2.15: Number of Assessment Units by Category for Uttar Pradesh, India. Source: National compilation on dynamic ground water resources of India, 2017, CGWB (CGWB, 2019)

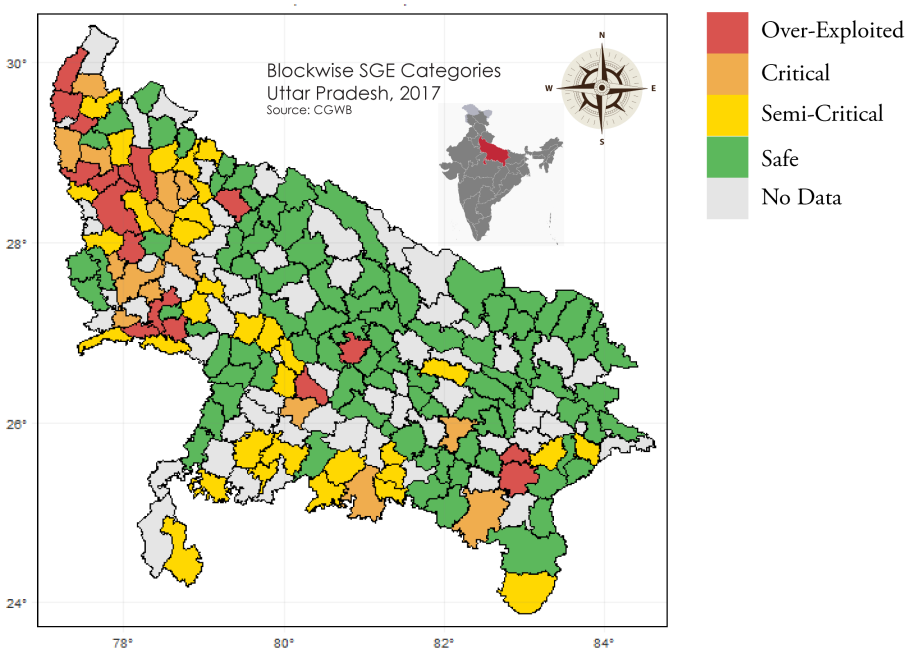


Figure 2.11: Categorization of Assessment Units in Uttar Pradesh. Source: National compilation on dynamic ground water resources of India, 2017, CGWB (CGWB, 2019).

Appendix

2.A Unconditional Mean Estimation Bias

Claim: If in a partially observed groundwater data sample G containing n observations at wells that each have the same maximum depth $\bar{g}$, d observations ($0 < d < n$) are missing due to dry wells and $\tilde{n}$ observations ($0 \leq \tilde{n} < n$) are missing due to causes other than dry wells (which we denote "other-causes" or "oc"), then $\hat{\mu}_{obs}$ is a biased estimator of the mean μ and the magnitude of bias is directly proportional to the fraction of dry wells $\frac{d}{n}$.

Proof:

Let $\hat{\mu}_{full}$ be the sample mean of the full data and $\hat{\mu}_{obs}$ be the sample mean of only the observed data. Let $\mu = E[G]$ be the population mean of the groundwater level so by definition of the sample mean, $E[\hat{\mu}_{full}] = \mu$. In the case of full data, the natural estimator for μ is the sample mean $\hat{\mu}_{full}$ given by equation 2.31.

$$\hat{\mu}_{full} = \frac{\sum_{i=1}^{n} g_i}{n} = \frac{\sum_{m_i=0} g_i + \sum_{m_i=1,dry} g_i + \sum_{m_i=1,oc} g_i}{n} \tag{2.31}$$

Where the last two summations are respectively the sums of the missing values due to dry wells and the missing values due to other causes. In general both types of missing observations (due to dry wells and due to other causes) are unknown so we cannot compute the full sample mean directly according to the formula in 2.31. We can, however, compute a lower bound for $\hat{\mu}_{full}$ since we know the lower bound of the GWL in dry wells.

Assertion: If well i is dry, then we know for sure that the GWL g_i is some value

that is greater than the maximum depth of the well, i.e, $g_i \geq \bar{g}$. This follows from the property that a well is said to "go dry" only after the water level falls below its floor which is its maximum depth.

We use the property of a dry well to compute a lower bound of full sample mean which call a "lower bound of the full sample mean" and we denote it by $\hat{\mu}_{lb}$. We compute the lower bound of the full sample mean by making the substituion $g_i = \bar{g}$ for all wells known to be dry. Making the substitution and knowing that there are d dry wells, we get the following result:

$$min(\sum_{m_i=1,dry} g_i) = (d \times \bar{g}) \tag{2.32}$$

Using 2.31 and 2.32, we may write the following expression for $\hat{\mu}_{lb}$:

$$\hat{\mu}_{lb} = \frac{\sum_{m_i=0} g_i + (d \times \bar{g}) + \sum_{m_i 1=1,oc} g_i}{n} \tag{2.33}$$

The sample mean of observed values μ_{obs} is given by equation 2.34.

$$\hat{\mu}_{obs} = \frac{\sum_{m_i=0}^{n} g_i}{n - d - \tilde{n}} \tag{2.34}$$

Rearrange to get the following expression:

$$\sum_{m_i=0} g_i = \hat{\mu}_{obs}(n - \tilde{n} - d) \tag{2.35}$$

We may now rewrite the expression 2.33 for $\hat{\mu}_{lb}$ using 2.35 as follows:

$$\hat{\mu}_{lb} = \frac{\hat{\mu}_{obs}(n - \tilde{n} - d) + (d \times \bar{g}) + \sum_{m_i=1,oc} g_i}{n} \tag{2.36}$$

Rearrange,

$$\hat{\mu}_{obs} = \frac{n\hat{\mu}_{lb} - (d \times \bar{g}) - \sum_{m_i=1,oc} g_i}{(n - \tilde{n} - d)} \tag{2.37}$$

Take expectations on both sides,

$$E[\hat{\mu}_{obs}] = \frac{nE[\hat{\mu}_{lb}] - (d \times \bar{g}) - \sum_{m_i=1,oc} E[g]_i}{(n - \tilde{n} - d)} \tag{2.38}$$

Let us assume that well observations that were missing due to oc wells were missing completely at random, so $E[g_i] = \mu$, thus we have, $\sum_{m_i=1,oc} E[g]_i = \tilde{n}\mu$ and $\bar{g},\ n,\ d$ are constants.

Substituting, we get,

$$E[\hat{\mu}_{obs}] = \frac{nE[\hat{\mu}_{lb}] - d\bar{g} - \tilde{n}\mu}{(n - \tilde{n} - d)} \neq \mu,\ in\ general \tag{2.39}$$

Thus in the presence of dry wells, $\hat{\mu}_{obs}$ is biased.

If there is no missingness due to other causes, or equivalently, if dry wells are the only cause of missingness then we have $\tilde{n} = 0$. We can simplify 2.36 to the following:

$$\hat{\mu}_{lb} = \frac{(n-d)\hat{\mu}_{obs} + d\bar{g}}{n} = (1 - \frac{d}{n})\hat{\mu}_{obs} + \frac{d}{n}\bar{g} \tag{2.40}$$

Equation 2.40 can also be written as follows:

$$\hat{\mu}_{lb} - \hat{\mu}_{obs} = \frac{d}{n}(\bar{g} - \hat{\mu}_{obs}) = b_{min} \tag{2.41}$$

Since $min(\hat{\mu}_{full}) = \hat{\mu}_{lb}$, the expression in 2.41 gives us the minimum difference between the truncated mean $\hat{\mu}_{obs}$ and the full sample mean $\hat{\mu}_{full}$. This difference that

we denote as b_{min} is an indicator of the minimum bias that occurs on account of missing values due to dry wells. The bias in the estimation of the unconditional spatial mean is thus a function of the fraction of dry wells, $\frac{d}{n}$, the truncated sample mean and the maximum well depth. If dry wells are very few, i.e, for $n >> d$, we have $\frac{d}{n} \approx 0$ so the RHS is close to 0 and we have $\hat{\mu}_{lb} = \hat{\mu}_{obs} \approx \hat{\mu}_{full}$.

2.B Monte Carlo Simulation: OLS in the presence of dry wells

Recall that we want to estimate the causal impact of groundwater recharge due to rainfall on GWLs using a simple linear regression of G on R. We have posited a true model given by the following equation:

$$G = R\beta + u \tag{2.42}$$

where the total monsoon rainfall received in this year in a circular region around a monitoring well is the covariate R (mm), the dependent variable is the post-monsoon GWL at that well G (mbgl), $u \sim N(0, \sigma^2)$ is an identically, normally distributed error and each error is independent with a mean 0 and constant variance σ^2. The parameter vector β is a 2×1 vector containing the constant of the regression β_0 and the slope of rainfall β_1. We want to estimate the parameters β from a sample of n observations $\{\boldsymbol{g}_{n\times 1}, \boldsymbol{r}_{n\times 1}\}$ where the $r_i's$ are fully observed but some of the $g_i's$ maybe dry and therefore missing as indicated by the column vector of binary missingness indicators $\vec{\boldsymbol{m}}_{n\times 1}$.

We employ a simulation sample size n = 8000. For our GWLs dataset, a cross-section for every year has close to 8000 observations. Cross-sectional data is often used for policy making especially during the monitoring years in which administrative agencies conduct national groundwater assessments (CGWB, 2019). We thus choose a sample size that best represents a policy making scenario. For comparison, a recent nationwide study on groundwater in India uses cross-sections of GWL data with sample sizes 1444, 1748 and 3192 spread over the whole country (Bhattarai et al., 2023). Our

sample of 8000 for one state is thus sufficient and still not excessively large. Note that the state level monitoring network for U.P (our study region) is more dense, i.e it has more well locations per unit area, than the nationwide network managed by the central government (Government of Uttar Pradesh, n.d.; CGWB, 2019).

The relationship between rainfall and GWL is a complex dynamic that depends on aquifer characteristics, geography, land-use patterns, soil quality, temperature and vegetation index. Numerous regional studies in hydrology have reported plausible linear relationships between rainfall and observed GWLs with R^2 ranging from 0.83 to 0.96 (Hussain et al., 2022) while non-linear relationships have also been explored (Cobb & Harvey, 2019). As an initial foray in studying the impact of missing data, we restrict ourselves to the linear case.

We assume that rainfall is exogenous to the GWL and by theoretical consequence, that the regression given by equation 2.42 captures a directional causality from right to left quantified by the estimated value $\hat{\beta}_1$. Since rainfall contributes to groundwater recharge, areas of relatively higher rainfall typically have shallower groundwater levels. This means that an increase in the numerical magnitude of rainfall will tend to cause a decrease in the magnitude of GWL mediated through the process of aquifer recharge. We thus assume that rainfall in a spatial neighborhood of a well is inversely correlated with the GWL at that well by the correlation coefficient $\rho_{RG} < 0$.

We discuss in detail specific instances of the simulation where the value of ρ_{RG} is low-negative (-0.26), high-negative (-0.85), low-positive (0.26) and high positive (0.85). We also study and report the impact on the OLS estimator for values of ρ_{RG} varying from -0.9 to 0.9 by increments of 0.1. This range includes the values obtained from the data in our study that lie between -0.14 and -0.26. An inverse relationship between GWL and rainfall implies that GWL are shallower (lower magnitude of depth) in areas where rainfall is high and vice versa. We therefore expect the true value of the slope parameter of lagged rainfall β_1 and the correlation ρ_{RG} to be strictly less than zero.

Denote the OLS estimator obtained with the regression performed on the full data as $\hat{\beta}_{full}$ and that obtained with only the complete cases as $\hat{\beta}_{obs}$. We simulate n observations for rainfall R and groundwater level G from a joint gamma distribution with mean $k = \{k_R, k_G\}$ and dispersion ($\theta = \{\theta_R, \theta_G\}$ where the correlation between R and G is

ρ_{RG}.

The choice of gamma distribution to model GWL and rainfall data is due to two reasons. First, both GWL and rainfall are non-negative quantities so a normal distribution assumption would not be appropriate considering that this assumption will likely produce inconsistent estimates. Second, a canonical text on missing data theory by Little and Rubin (2019) employs the exponential distribution to provide analytical results for missing data in case of a random variable censored beyond a fixed point. This case is analogous to our GWL variable that is censored beyond the maximum well depth $\bar{g}$ (Little & Rubin, 2019, p.136-137). The exponential distribution is a special case of the gamma family of distribution. Hence, the choice of the gamma distribution ensures coherence with the existing literature while also generalizing the approach and findings.

In practice, OLS does not require normally distributed errors for unbiasedness and consistency but only for hypothesis testing and obtaining confidence intervals (Wooldridge, 2013, p.59). For our purpose, which is to perform bias estimation due to missing data and construct confidence intervals via Monte Carlo simulation, we may proceed with a Gamma distribution assumption without any impact on results.

The simulated n observations constitute the full data. For each observation we obtain a realization of GWL g_i and rainfall r_i. If the realization g_i is greater than the screening depth $\bar{g}$, then well i is dry. The number of GWL values greater than the predetermined screening depth $\bar{g}$ is denoted d; it is the number of missing values due to dry wells. The set of observed values contains only the $n - d$ complete cases where $g_i \leq \bar{g}$. We also obtain from the simulated data a missingness indicator M of dimension $n \times 1$ s.t $m_i = 1$ if g_i is missing and 0 otherwise. We then regress groundwater level on rainfall as given by equation 2.42 using the full data to get $\hat{\beta}_{full}$ and then with the observed values only to get $\hat{\beta}_{obs}$.

The simulation, truncation and regressions constitute one iteration of the Monte Carlo simulation. We repeat this procedure N_{sim} times and the mean of the OLS estimators computed over these iterations constitutes the final estimated values.

Table 2.B.1 lists and defines the inputs (true parameters, sample size, number of iterations) and outputs (simulated data, missingness indicator, number of dry wells and

Parameter	Dimension	Description	Value
Inputs			
n	Scalar	Sample size full data	8000
$k = \{k_R, k_G\}'$	2×1	Mean of bivariate gamma distr.	$\{900, 9.75\}'$
$\theta = \{\theta_R, \theta_G\}'$	2×1	Dispersion of bivariate gamma distr.	$\{0.03, 0.39\}'$
ρ_{RG}	Scalar	Correlation coeff., rainfall and GWL	[-0.9,0.9]
N_{sim}	Scalar	Number of simulation iterations	100
$\bar{g}$	Scalar	Maximum depth of well	99th percentile GWL
Outputs			
G	$n \times 1$	Random vector; outcomes (gw levels)	
R	$n \times 1$	Random vector; covariates (rainfall)	
M	$n \times 1$	Random binary vector; (missingness)	
d	Scalar	Number of missing values (dry wells)	
$\hat{\beta}_{1,full}$	Scalar	Estimated slope(rainfall), full data	
$\hat{\beta}_{1,obs}$	Scalar	Estimated slope(rainfall), obs. data	
$Var[\hat{\beta}_{1,full}]$	Scalar	Var. of OLS estimator, full data	
$Var[\hat{\beta}_{1,obs}]$	Scalar	Var. of OLS estimator, obs. data	

Table 2.B.1: Inputs and outputs of the Monte Carlo simulation study.

OLS estimates) in the simulation exercise. The true parameters consist of the mean k and dispersion θ of a bivariate gamma distribution. The first variable in this gamma distribution is the rainfall R and the second is the GWL G. Proxies for the true parameter values were obtained by first fitting gamma distributions separately to the real-world rainfall and GWLs data. In particular, the estimated values of mean and dispersion that maximized the likelihood of observing each real-world sample were taken to be the "true" parameters. Table 2.B.2 shows the summary statistics of the real-world data along with those of the simulated data generated by using the true parameter values obtained via 100 Monte Carlo iterations. The distributions of simulated data closely match the real-world data as expected.

The purpose of Monte Carlo iteration in our simulation exercise is to construct confidence intervals for the estimated bias. With 100 iterations, we obtain estimates with very small confidence intervals (with size of the order of 0.0001, see Figure 2.B.2). Increasing the number of iterations could reduce the size of the intervals further. This may be statistically significant although it should not have any impact in practice as changes at the fourth place of decimal will likely not influence policy outcomes.

Statistic	N	Mean	St. Dev.	Min	Max
Simulated Rainfall	8,000	899.9	155.7	430.02	1,619.6
Real-world Rainfall	7,954	955.2	199.3	563.9	1659.5
Simulated GWL	8,000	9.76	6.1	0.16	50.9
Real-world GWL	7,954	9.75	6.9	0.15	53.52

Table 2.B.2: Summary statistics of real-world and simulated data.

2.B.1 Simulation Results

These results are primarily presented in the results section 4.4.2. They are replicated here from section 4.4.2 for easy access and readability. Figure 2.B.1 shows four different configurations of the Monte Carlo simulation. They are scatter plots of simulated, jointly gamma-distributed GWL and rainfall observations in scenarios of negative (left top and bottom panels) and positive (right top and bottom panels) correlations between rainfall and GWL. The top-left panel shows the realized outcomes when the correlation coefficient ρ_{RG} is set to -0.26 which is a scenario of low negative correlation. The bottom-left panel shows the outcomes for $\rho_{RG} = -0.85$ or high negative correlation. Similarly, the right top and bottom panels show the outcome for $\rho_{RG} = 0.26$ and $\rho_{RG} = 0.85$ respectively.

We have already shown analytically in section 2.3.5 that bias will be higher when the covariance between GWL and rainfall is higher. We recall the pertinent analytical result here. If $\bar{r}_{full}$ and $\bar{r}_{obs}$ are the sample means of the rainfall observations with the full and observed data respectively and similarly $\bar{g}_{full}$ and $\bar{g}_{obs}$ for GWLs and for observations $i = 1, ..., r*$ the GWLs are observed and for $i = r* + 1, ..., N$ they are missing, then we have for the full data:

$$\hat{\beta}_{1,full} = \hat{\beta}_{1,obs}.k + \frac{(\bar{r}_{obs} - \bar{r}_{full})\bar{g}_{obs}}{\hat{\sigma}^2_{r,full}} + \frac{\sum_{i=r*+1}^{n} (r_i - \bar{r}_{full})g_i}{\hat{\sigma}^2_{r,full}} \tag{2.43}$$

where $\hat{\sigma}^2_{r,full}$ is the variance of rainfall in the full data and $\hat{\sigma}^2_{r,obs}$ is the variance of the observed data and $k = \frac{\hat{\sigma}^2_{r,obs}}{\hat{\sigma}^2_{r,full}}$. Subtract $\hat{\beta}_{1,obs}$ from both sides to obtain an expression for the bias as follows:

$$\hat{\beta}_{1,full} - \hat{\beta}_{1,obs} = \hat{\beta}_{1,obs}.(k-1) + \frac{(\bar{r}_{obs} - \bar{r}_{full})\bar{g}_{obs}}{\hat{\sigma}^2_{r,full}} + \frac{\sum_{i=r*+1}^{n} (r_i - \bar{r}_{full})g_i}{\hat{\sigma}^2_{r,full}} \quad (2.44)$$

Equation 2.44 has three terms on the RHS that determine the size and direction of bias. The value of the first term is determined by the value of k, the ratio of the sample variance of truncated rainfall data to that of the full rainfall data. If the two variances are roughly equal then $k \approx 1$ and this term is close to 0 and bias is lowered. The second and third terms are both scaled by the reciprocal of the full sample rainfall variance which is a constant so we may only focus on the numerators. The size of the second term is primarily determined by the difference in the means of the truncated and full sample rainfall data. If the two means are roughly equal, this term is close to 0 and bias is lowered. The third term pertains only to the truncated portion of the rainfall data. If the covariance between rainfall and GWL *within the truncated portion* of data is high, bias is higher and lower otherwise. We summarize our analytical assertions below:

1. If the sample variance of the truncated rainfall data is significantly different from that of the full rainfall data, i.e if k is either close to 0 or much greater than 1, bias will tend to increase.
2. If the sample mean of the truncated rainfall data is significantly different from that of the full rainfall data, i.e if $|(\bar{r}_{obs} - \bar{r}_{full})|$ (see the second term of equation 2.44) is significantly greater than 0, bias will tend to increase.
3. If the covariance between GWL and rainfall in the truncated portion of data is higher, bias will tend to increase.

The realized outcomes shown in Figure 2.B.1 corroborate these analytical assertions and depict them visually. Bias when correlation is either 0.85 or -0.85 is 0.002, twice as much as 0.001 which is the bias for correlations of 0.26 and -0.26. This suggests that the direction of correlation, either negative or positive, does not matter.

Recall that when we listwise delete missing data due to dry wells, we lose the GWL observations as well as the corresponding rainfall observations. If we focus on the truncated portion of data shown as red dots in the Figure 2.B.1, we see that in cases

of high correlation, the data that is lost to listwise deletion contains more "signal" or information as compared to that when correlation is low. Table 2.B.3 describes the effect of the four types of correlation (low-negative, high-negative, low-positive, and high-positive) on the three different terms in equation 2.44.

When correlation of GWL with rainfall is either low-negative or low-positive, dry wells are likely to be "scattered" across the whole range of rainfall values as seen in the top-left and bottom-left panels of Figure 2.B.1. Deleting these values does not reduce the sample variance of the rainfall data and thus $k \approx 1$ and the first term is close to 0. Further, when correlation is low-negative or low-positive, then the truncation of data that is scattered across the distribution of rainfall does not significantly change the sample mean and thus the second term is also close to 0. With the first two terms tending to zero, for low correlation, the direction of the bias is mostly determined by the third term which matches the sign of the correlation. This is visible in Figure 2.B.1 (top-left and bottom-left panels) where the bias is negative for both low-negative and low-positive correlation. The truncated OLS estimate in both cases reveals a "weaker" relationship than the full sample OLS estimate.

Conversely, when correlation is high-negative, dry wells tend to occur in areas of low rainfall and truncation causes the left tail of the rainfall distribution to be lost. This reduces the sample variance of the truncated rainfall data compared to the full rainfall data and thus $k < 1$ and the first term is negative. Further, if correlation is high-negative, the sample mean of the truncated data is higher than the sample mean of the full data and the second term is positive. The third term measures covariance within the truncated data which has the same sign as correlation, negative. Thus, in the case of high-negative correlation, the first and third terms are negative while the second is positive. The simulation reveals that for the case we consider, the negative sign dominates and the overall bias is negative.

If the correlation is high-positive, dry wells occur in areas of high rainfall (a physically improbable situation) and we lose the right tail of the rainfall distribution and thus $k < 1$ and the first term is still negative. The sample mean of truncated rainfall is lower than that of the full sample data and the second term is negative. The third term is, like correlation, positive. The simulation reveals that the negative sign dominates and the

Term	ρ_{RG}	Effect	Sign	Bias
First	Low negative	$k \approx 1$	0	≈ 0
Second	Low negative	$\lvert(\bar{r}_{obs} - \bar{r}_{full})\rvert \approx 0$	0	≈ 0
Third	Low negative	$\sum_{i=r*+1}^{n}(r_i - \bar{r}_{full})g_i < 0$	-ve	< 0
First	High negative	$k < 1$	-ve	< 0
Second	High negative	$(\bar{r}_{obs} - \bar{r}_{full}) > 0$	+ve	> 0
Third	High negative	$\sum_{i=r*+1}^{n}(r_i - \bar{r}_{full})g_i < 0$	-ve	< 0
First	Low positive	$k \approx 0$	≈ 0	< 0
Second	Low positive	$(\bar{r}_{obs} - \bar{r}_{full}) \approx 0$	0	≈ 0
Third	Low positive	$\sum_{i=r*+1}^{n}(r_i - \bar{r}_{full})g_i > 0$	+ve	> 0
First	High positive	$k < 1$	-ve	< 0
Second	High positive	$(\bar{r}_{obs} - \bar{r}_{full}) < 0$	-ve	< 0
Third	High positive	$\sum_{i=r*+1}^{n}(r_i - \bar{r}_{full})g_i > 0$	+ve	> 0

Table 2.B.3: Correlation between GWL and Rainfall and Expected OLS Estimation Bias

overall bias is negative.

Figure 2.B.2 shows the results of a 100 iterations of a Monte Carlo simulation to study the impact of correlation on the mean (top-left panel) and variance (bottom-left panel) of the OLS estimators, the truncation bias (top-right panel) and the root-mean-squared-error (RMSE) (bottom-right panel) of the regression. The interesting feature of these results is that the variance of the OLS estimators and the RMSE are greater for the full data as compared to the truncated data. Further, the difference in variance and RMSE is greater when the correlation is lower and it decreases in either the negative or positive direction. This means that for the low correlation case, the OLS estimator is biased downwards and has lower variance than the OLS estimator with full sample data. From the graphs shown in Figure 2.B.2, we can see that in the case of low correlation, adding the red dots into the the truncated sample (gray dots) actually *adds* some uncertainty whereas in the case of high correlation the red dots adds information that strengthens the existing relationship.

Listwise deletion of missing data thus poses different hazards for causal inference in low and high correlation scenarios. For low correlation scenarios, a weaker relationship

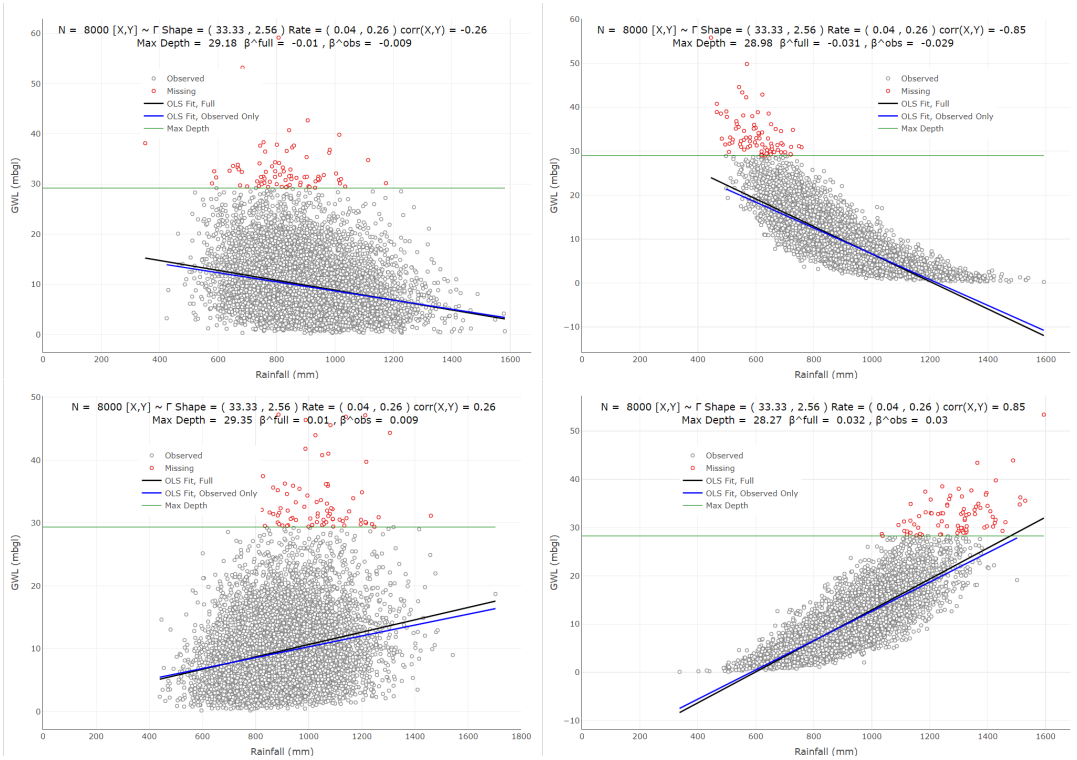


Figure 2.B.1: OLS with Listwise Deletion of Simulated Missing Data due to Dry Wells: The Impact of Correlation Between GWL and Rainfall. The cases of low negative correlation (top-left panel), high negative correlation (top-right pabnel), low positive correlation (bottom-left panle) and high positive correlation (bottom-right panel).

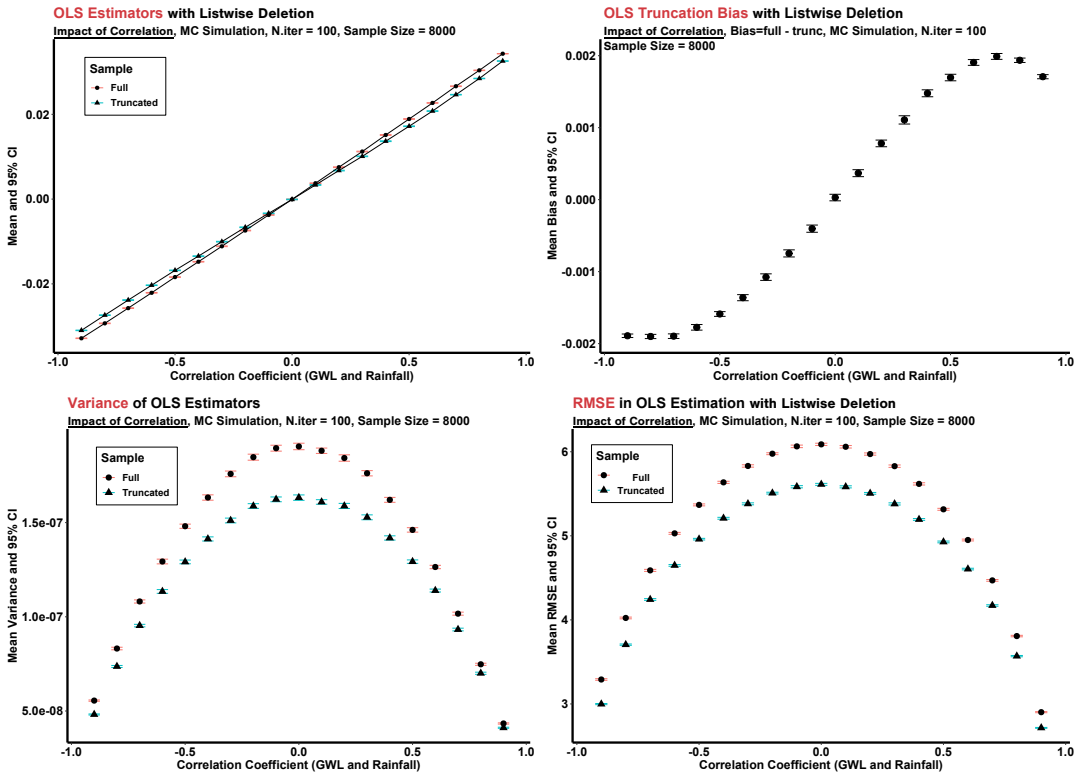


Figure 2.B.2: OLS with Listwise Deletion of Missing Data due to Dry Wells: The Impact of Correlation Between GWL and Rainfall

of GWL with rainfall will be inferred with greater certainty but the extent of bias will be small (as in it won't appear too much weaker). For high correlation scenarios, there won't be a great effect on the variance of OLS estimators but a significantly weaker relationship between GWL and rainfall will be inferred than if we were to estimate with full data.

These findings are pertinent to policy especially for estimating the impacts of drought. The truncation bias resulting from dry wells can lead to underestimation of the causal impact of drought on groundwater levels and thus on irrigation water availability. For example, consider the result obtained for high negative correlation (-0.85, see Figure 2.B.1, top-right panel) with truncated data is $\hat{\beta}_{1,obs} = -0.029$. This implies that on average, a decrease in rainfall by 1 mm will cause the deepening of GWL by 29 mm (or 0.029 meters). Now, the result obtained with the full data is $\hat{\beta}_{1,full} = -0.031$ which means that GWLs will deepen by 31 mm (0.031 meters). Truncation thus causes an underestimation of GWL decline by 2 mm or 6.5% of the true value (31 mm). The incidence and intensity of drought is likely to increase with climate change and one of the keys to mitigating the loss due to drought is the proper understanding of the relationship between rainfall and groundwater recharge. We have shown through this illustrative simulation that deletion of missing data will lead to *false optimism* in predicting the impact of drought on GWLs in the sense that the actual impact may be greater than that predicted by using truncated data (Arora & Ali, 2022).

2.C Simple Linear Regression with Truncated Data

Let $\hat{\beta}_{1,full}$ be the OLS estimate of the coefficient of rainfall X in the simple regression of equation 2.25 when the complete data is used and $\hat{\beta}_{1,obs}$ be the estimate when observations are truncated where groundwater level G is not observed. Let $\bar{x}_{full}$ and $\bar{x}_{obs}$ be the sample means of the rainfall observations with the full and truncated data respectively and similarly $\bar{g}_{full}$ and $\bar{g}_{obs}$ for groundwater levels. As before, assume that for observations $i = 1, ..., r$ the groundwater levels are observed and for $i = r+1, ..., n$ they are missing. Then we have for the full data:

$$\hat{\beta}_{1,full} = \frac{\sum_{i=1}^{n}(x_i - \bar{x}_{full})g_i}{\sum_{i=1}^{n}(x_i - \bar{x}_{full})^2} \tag{2.45}$$

and for the truncated data:

$$\hat{\beta}_{1,trunc} = \frac{\sum_{i=1}^{r}(x_i - \bar{x}_{obs})g_i}{\sum_{i=1}^{r}(x_i - \bar{x}_{obs})^2} \tag{2.46}$$

If $\hat{\sigma}^2_{X,full}$ and $\hat{\sigma}^2_{X,trunc}$ are the sample variances of the full and truncated rainfall data then we can write equation 2.45 as follows:

$$\hat{\beta}_{1,full} = \frac{\sum_{i=1}^{n}(x_i - \bar{x}_{full})g_i}{\sum_{i=1}^{n}(x_i - \bar{x}_{full})^2} \times \frac{\hat{\sigma}^2_{X,trunc}}{\hat{\sigma}^2_{X,trunc}} \tag{2.47}$$

$$= \frac{\sum_{i=1}^{n}(x_i - \bar{x}_{full})g_i}{\sum_{i=1}^{r}(x_i - \bar{x}_{obs})^2} \times \frac{\hat{\sigma}^2_{X,trunc}}{\hat{\sigma}^2_{X,full}} \tag{2.48}$$

Let $\frac{\hat{\sigma}^2_{X,trunc}}{\hat{\sigma}^2_{X,full}} = k$, then:

$$= \frac{k\sum_{i=1}^{r}(x_i - \bar{x}_{full})g_i + k\sum_{i=r+1}^{n}(x_i - \bar{x}_{full})g_i}{\sum_{i=1}^{r}(x_i - \bar{x}_{obs})^2} \tag{2.49}$$

Let $\bar{x}_{obs} = \bar{x}_{full} + p$, then substitute for $\bar{x}_{full}$ and simplify to get:

$$= \hat{\beta}_{1,trunc}.k + \frac{p\bar{g}_{obs}}{\hat{\sigma}^2_{X,full}} + \frac{\sum_{i=r+1}^{n}(x_i - \bar{x}_{full})g_i}{\hat{\sigma}^2_{X,full}} \tag{2.50}$$

Note that we obtain from equation 2.50 the conditions under which OLS estimators from full ($\hat{\beta}_{1,full}$) and truncated ($\hat{\beta}_{1,obs}$) data are equal. These conditions, after some simplification are as follows:

$$k = \frac{\hat{\sigma}^2_{X,trunc}}{\hat{\sigma}^2_{X,full}} = 1 \tag{2.51}$$

$$p\bar{g}_{obs} = (\bar{x}_{obs} - \bar{x}_{full})\bar{g}_{obs} = -\sum_{i=r+1}^{n}(x_i - \bar{x}_{full})g_i \tag{2.52}$$

The first condition (equation 2.51) is satisfied when the variance of the full and truncated rainfall data are equal. In other words, when removal of values from the full data does not reduce the variance of the sample.

2.D Data Construction

If a groundwater level observation i was made in year t at well location (x, y) then we define a spatio-temporal neighborhood of size (r, l) of this observation as

$$N_{s,i} \;=\; \{\, (x', y', t') :\; s.t\, |(x', y') - (x, y)| \leq r \; and \; t' = t - l \,\} \tag{2.53}$$

where r is called the "radius" and l the "lag" of the neighborhood. Note that the lagged time period t' can only be in the past thus we must have $l > 0$ and $t' < t$. If the lag is zero, then we refer to the observation in the current monitoring year. We use the concept of the spatio-temporal neighborhood to define a set of spatial variables for climatic parameters and land-use categories for each groundwater level observation in the data. To obtain a value for a climatic or land-use variable for a given groundwater

level observation, we extract data from the remotely sensed raster in a spatio-temporal neighborhood of the well and year at which the groundwater level observation was made.

We use gridded geo-spatial time-series climate data from the ERA5 land data set available as a stack of raster grids at a resolution of $0.08^\circ \times 0.08^\circ$ (approximately 9km $\times$ 9 km) through the Copernicus Climate Data Store (see Figure 2.D.3) (Muñoz Sabater, 2019). Each layer in the raster stack contains monthly averages for 322 months starting in January 1994 up to October 2020 for "total precipitation" and "2 meter temperature." Using the stack of raster imagery, we construct a spatial variable that measures total rainfall in the monsoon months (June, July, August, September) in the year of monitoring and the year immediately prior to the year of monitoring. We also extract mean temperature during the month of May in the year of monitoring. Figure 2.D.3 shows spatial plots of the average monthly mean of lagged total monsoon precipitation.

To capture the value of a climatic variable in a spatio-temporal neighborhood, we take its value at the nearest grid point to the well location in question lagged by the lag value. Figure 2.D.1 shows the construction procedure. The grid resolution enforces a minimum spatial neighborhood radius of approximately 4.5 km thus for the climatic variables, the value for any particular well and year of monitoring reflects the aggregated value in a radius of 4.5 km around that well. Table 2.D.1 provides a summary of the constructed spatial variables.

We acquired categorical LULC data through the public portal of the Indian Space Research Organization (ISRO) for UP from 2005-06 to 2014-15 (ISRO, 2022). The data is a series of gridded raster images containing LULC categories for each grid cell at a spatial resolution of 56 meters. This means that each grid cell occupies an area of 56m $\times$ 56m (approximately 0.003 sq km.) on the ground (NRSC, 2007). Figure 2.D.4 shows maps of land-use land cover for the Indo-Gangetic Plain as well as a magnified map for the state of UP.

We construct a set of variables that capture percentage area by land-use category to characterize land-use around each groundwater level observation within approximately a 1 km radius of the observation well in the current monitoring year and lagged by one year in the past. We construct a variable each for percentage area under double-triple

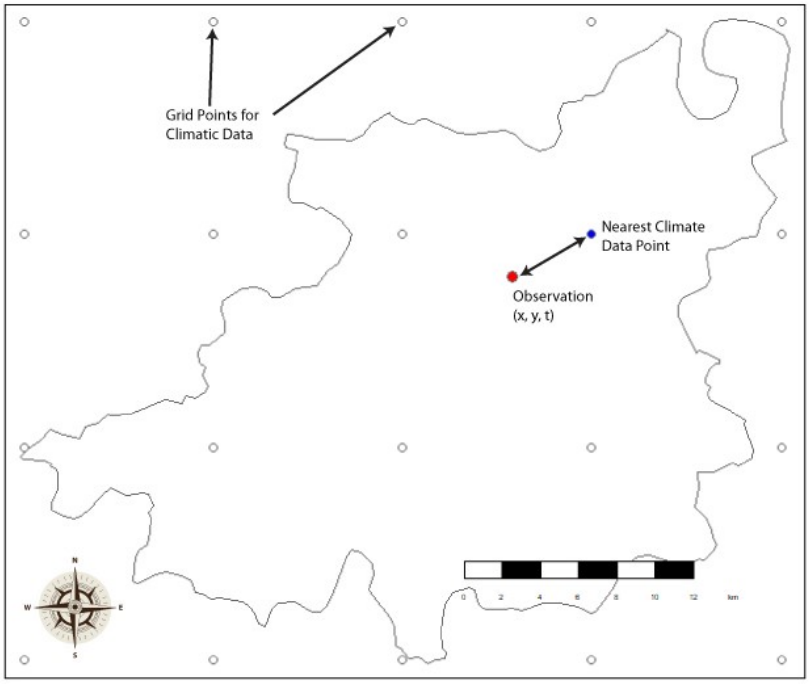


Figure 2.D.1: Construction of climatic spatial variables from gridded satellite data.

Figure 2.D.2: Land-Use in a spatio-temporal neighborhood of a groundwater level observation.

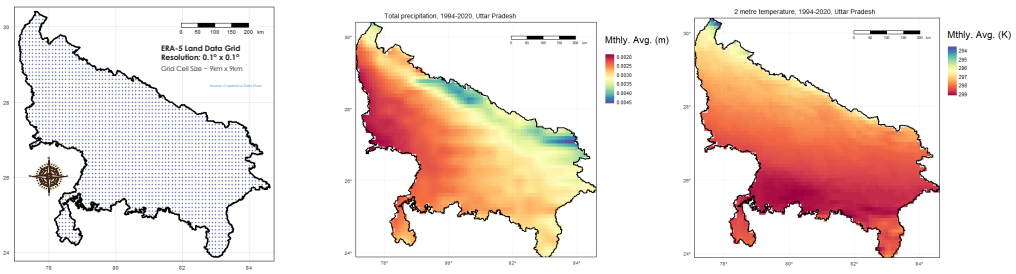


Figure 2.D.3: Climatic Variables Spatial Means: Data Grid and Spatial Mean of Total Monsoon Precipitation and Monthly Avg. Temperature, Uttar Pradesh, 1994-2020.

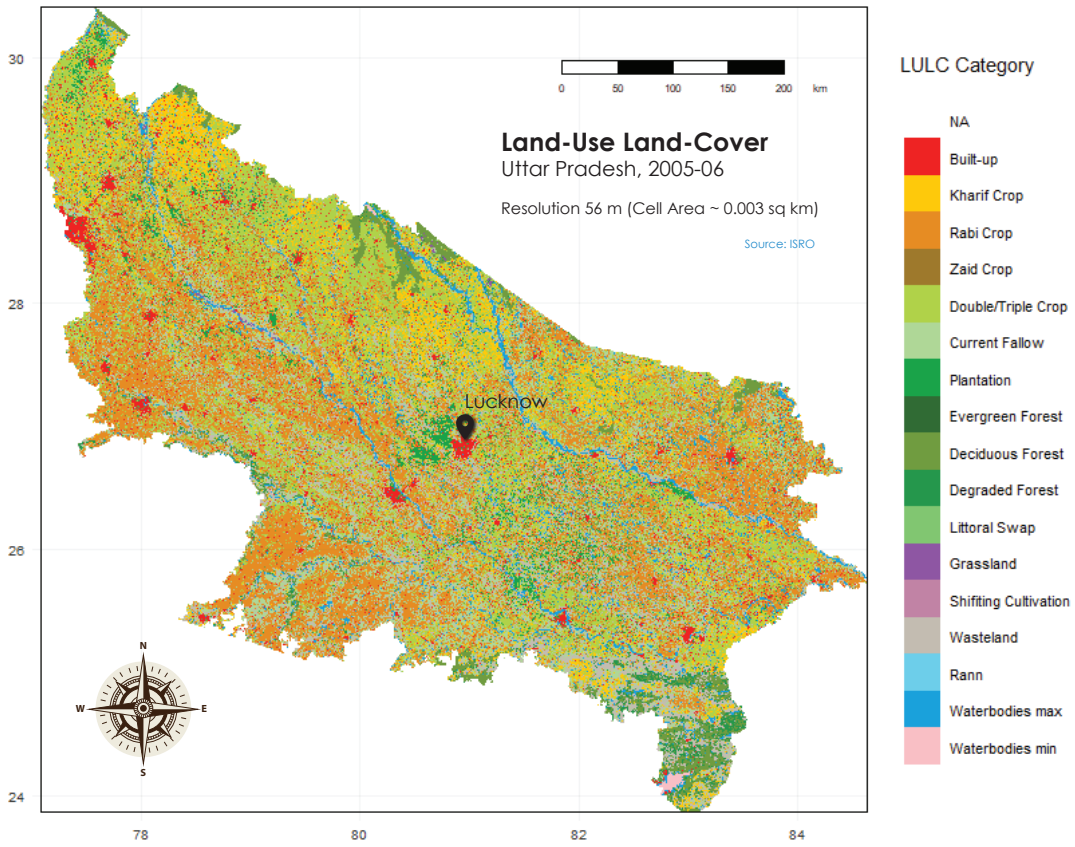


Figure 2.D.4: Categorical Land-Use Land-Cover Map of Uttar Pradesh, 2005-06. Each grid cell or "pixel" is assigned a category (coded by colors in the legend) that describes the land use or land cover situation in the given year. The data consists of 10 such images for the years from 2005-2015. Source: ISRO (ISRO, 2022)

Name	Description (Unit)	Lag	Range	Mean (SD)	% Missing
L0RF	Monsoon Rfl. (m)	0	0.29-2.3	0.88 (0.23)	0
L1RF	Lagged Monsoon Rfl. (m)	1	0.29-2.3	0.91 (0.27)	0
L0T2M	Mean May 2m Temp. (K)	0	270-309	306 (1.27)	0
L1PBUILTUP	Lagged Built-Up Area (%)	1	0-1	0.07 (0.11)	0
L1PDTC	Lagged Double/Triple Crop Area (%)	1	0-1	0.52 (0.25)	0
L1PRABIZAID	Lagged Rabi/Zaid Area (%)	1	0-1	0.13 (0.18)	0
L1PKHARIF	Lagged Kharif Area (%)	1	0-0.9	0.07 (0.10)	0

Table 2.D.1: Constructed Spatial Variables. Lagged Rainfall (L1RF) is available from 2009 to 2019 and the others are available between 2009-10 and 2014-15.

cropping, rabi-zaid (dry-season) cropping, kharif (wet-season) cropping and built-up settlement in the previous year. For example, for a groundwater level observation at well located at (x, y) in year t, we compute land-use in a neighborhood of size r around (x, y) in year $t - 1$. Lags of 1 year are indicated by the prefix $L1$ in the variable name. Land-use variables are only constructed for a subset of the study period for the years 2009-10 to 2014-15 as per the availability of the data. Figure 2.D.2 shows the construction of the spatial neighborhoods.

CHAPTER 3

SPATIAL AUTOCORRELATION IN GROUNDWATER LEVELS

3.1 Introduction

Groundwater contained in the porous soils and rocks of underground aquifers is not stationary, but has a definite flow due to pressure differentials between underground locations (USGS, 1902). The flow of water in underground aquifers from regions of high pressure to low pressure is called "sub-surface flow."

When a groundwater user pumps water through a well at one location, the removal of water forms a zone of relatively lower pressure[1] at the pumping location. This differential causes "inflow" of water toward the pumping well from nearby locations. The decline in groundwater level (GWL) at the pumping location[2] causes a decline in GWLs in the spatial neighborhood of the pumping well. Groundwater pumping thus creates a "spatial externality" whereby the extraction by one user influences the GWLs experienced by nearby users (Pfeiffer & Lin, 2012). Similarly, when water from rainfall infiltrates the ground surface and travels downward into an aquifer at one location, the resultant rise in GWLs is distributed to nearby locations through the "outflow" of water from the point of infiltration. The spatial dispersion of the effects of pumping and recharge in an aquifer gives a groundwater aquifer the nature of a non-exclusive, common pool resource (CPR) whereby the costs and benefits of resource usage are shared by spatially proximate users (Pfeiffer & Lin, 2012; Edwards, 2016).

According to Gardner, Ostrom and Walker (GOW) (Gardner et al., 1990; Ostrom, 1990), groundwater extraction by farmers and other users from a common aquifer is a case of "multiple interdependent appropriators" who draw from the same CPR. Governance of the CPR in the presence of multiple appropriators makes it necessary to

[1] Often called a "cone of depression"
[2] Often called the "drawdown"

delineate the size and boundaries of aquifer regions that constitute independent spatial "units" within which users share water resources. It is important both for the individual appropriators and the collective to know the spatial range within which pumping externalities occur as well as the zone of recharge that replenishes the aquifer. The identification of these zones is an active area of research in hydrology but the interaction of groundwater governance institutions and the zonal characteristics of spatially heterogeneous aquifers is relatively scarce (Edwards, 2016).

In this chapter, we explore the spatial characterization of aquifer structure through purely data-driven methods. Specifically, we exploit the structure of "spatial auto-correlation" in GWLs to extract relevant information about resource sharing in an aquifer. Spatial auto-correlation refers to the correlation between the observed value of the variable at one location to that of *its own* (auto) values at nearby locations (N. A. Cressie, 1993; Anselin & Bera, 1998). Our approach rests on the assumption that higher positive correlation between GWLs at two different locations implies a greater extent of resource sharing between users at those locations. We expect that standing at any given point in the aquifer, GWLs at nearby locations should be more similar to the GWL at our location than those that are further away. Spatial auto-correlation in GWLs must therefore decrease with distance. Our empirical results from a case study corroborate that such a structure of spatial autocorrelation is indeed present in an alluvial aquifer system in northern India.

We estimate spatially autocorrelation in GWLs at varying distances within the aquifer using a functional instrument called the "semivariogram." The "range" parameter of the semivariogram represents the distance within which spatial autocorrelation is "significantly" high. We consider this distance to be the size of a "spatial unit" in which the resource "acts as a singly entity" and can be said to belong to a shared pool amongst overlying users. Users separated by a distance greater than the range parameter are independent *vis-à-vis* resource usage.

Sub-surface flow of groundwater through the aquifer has the effect of locally "leveling out" any differentials in GWL created due to pumping and recharge events giving rise to spatial autocorrelation in GWLs. Due to the leveling-out effect of sub-surface flow spatial continuity in hydro-geological, climatic and socio-economic factors, GWLs

at locations close to each other tend to have similar magnitude as compared to those far-away hence they are usually found to exhibit *positive* spatial auto-correlation (Yan et al., 2006; Gundogdu & Guney, 2007; Moukana & Koike, 2008; Varouchakis et al., 2009; Varouchakis et al., 2012; Machiwal et al., 2012; Liu et al., 2013; Dong et al., 2019; Islam et al., 2021; Hasan et al., 2021)[3].

In a body of water like a bathtub where the movement of water is unmediated, the addition or removal of water from any point causes a uniform change in water-level across the whole tub. The bathtub thus acts as a *single unit* within which water level is perfectly spatially autocorrelated. In an aquifer, GWL dynamics are more complex since the movement of water is constrained by the aquifer media like alluvium, crystalline rocks, sand or gravel within which groundwater flows. The effects of pumping and recharge are therefore transferred only within some limited distance from a given point in the aquifer[4]. Further, the rate of sub-surface flow depends on spatially heterogeneous aquifer characteristics and the magnitude of the pressure differentials (USGS, 1902). A aquifer may consist of many spatial units within which GWLs exhibit spatial autocorrelation and these units may be connected to each other. An aquifer has been likened to an "egg-tray" which consists of many interconnected units instead of a single unit like a bathtub (Edwards, 2016).

The rate of groundwater flows through an aquifer medium is directly proportional to its "hydraulic conductivity" (USGS, 2013a). Hydraulic conductivity depends on the permeability and porosity of the soil and rocks that make up an aquifer. Permeability and porosity are themselves dependent on the size and distribution of interstitial spaces in the aquifer media. Sub-surface flow will be greater in areas where the hydraulic conductivity is high and the pumping externality and the dispersion of recharge water will correspondingly be greater in those areas. The water resource in these areas will be "shared" to a greater extent as compared to areas where hydraulic conductivity is smaller.

A question that has occupied water resource economists for a long time relates to

[3]By contrast, plant height in a garden is often reported to show negative spatial auto-correlation where if a plant grows taller at a faster rate it hoards up resources and tends to be surrounded by shorter plants (Griffith, 2019).

[4]The maximum distance from a pumping well at which the effects are "significant" is called the "radius of influence" of the pumping well (Bresciani, Shandilya, Kang, & Lee, 2020).

the differential welfare gains from optimally managed versus open-access groundwater extraction. Gisser and Sanchez (1980) proposed a theoretical model to compare welfare gains from groundwater management assuming a simple bathtub type aquifer. They found that for aquifers with "relatively large" water storage capacity, there are virtually no welfare gains (Gisser & Sanchez, 1980). This result depends critically on the assumption of the bathtub type of aquifer in which the effects of pumping at one location are immediately and equally transferred to all others. Brozovic *et. al* (2010) argue that the groundwater pumping externality has a "spatial dynamic nature" which when explicitly included in an economic model of groundwater management would yield results that could be orders of magnitude different from those coming out of a bathtub model, especially for large and complex aquifers (Brozović et al., 2010). Merrill and Guilfoos (2018) simulate the dynamics of depletion in a large aquifer and find that incorporating the spatial pattern of depletion in the aquifer increases the welfare estimates of moving from open access to optimal management (Merrill & Guilfoos, 2018).

Edwards (2016) has shown empirically that spatial heterogeneity in aquifer properties affects how welfare gains from resource management are distributed to groundwater users. A region overlying a part of an aquifer characterized by high hydraulic conductivity and low recharge may face a more costly common-pool problem and therefore benefit more from management (Edwards, 2016). The spatial characterization of aquifer properties is thus important to understand how groundwater users benefit from management.

In hydrology and related natural sciences, hydraulic conductivity and other hydraulic properties of aquifers are measured via physical methods. These include expensive field investigations like "pumping tests" where water is pumped at a well and GWL dynamics at that well and nearby wells are observed in response to pumping (Todd & Mays, 2005, p.100). These investigations are costly and required specialized equipment and expertise that is beyond the scope of economics. Besides, for groundwater management and policy-making, the actual results of pumping tests are not as important as the understanding of how they influence the common-pool nature of the aquifer. We propose an approach to characterize aquifer structure which only relies on GWL data circumventing the need for direct physical investigations of aquifer characteristics[5].

[5]Note that richer models could be developed if physical measurements were available. In India, the

Another important question in groundwater management is *where* management must be applied in order to gain optimal welfare gains. Users located over regions of the aquifer where groundwater resources are scarce or heavily depleted benefit the most from management (Gisser & Sanchez, 1980; Edwards, 2016). In GOW framework, this is a case of the "assignment problem" (Gardner et al., 1990). Users across all parts of the aquifer do not obtain the same welfare gains from management therefore it is necessary to spatially delineate how management regimes must be assigned to different regions in order to maximize gains.

Resource depletion in aquifers is typically indicated by high (deep) GWLs. GWLs are measured as the depth of water table from ground surface using monitoring wells. Monitoring wells produce point samples at different locations across the aquifer but to delineate regional boundaries, we need a continuous field of GWL measurements. We again rely on spatial autocorrelation to spatially predict GWLs at unsampled locations using existing GWLs point sample data available through governmental agencies. We use the "kriging" estimator which is the best unbiased estimator for linear spatial prediction (Krige, 1951; Matheron, 1971). It differs from other spatial interpolattion methods like inverse-distance weighting (IDW) because it also provides a prediction variance for each prediction (Bárdossy, 1997).

The spatial autocorrelation structure also holds importance for public policy evaluation and development economics as it pertains to attaining causal inference. If spatial auto-correlation is not accounted for, it may pose problems with identification and estimation (Beale, Lennon, Yearsley, Brewer, & Elston, 2010). Linear regression models have been previously employed to estimate the impact of irrigated agriculture, rainfall and policy interventions on regional GWL situation (i.e., GWL is a dependent variable) (Asoka et al., 2017; Bhanja et al., 2019) (Bhanja et al., 2017).

In a linear regression model estimated via ordinary least squares (OLS), spatial autocorrelation in the dependent variable will violate the classical assumptions of independent error terms (Anselin & Bera, 1998). Sub-surface flow is usually an *unobserved source of spatial auto-correlation* in GWLs. It is difficult to control for through the

Central Ground Water Board (CGWB) has undertaken a national aquifer mapping (NAQUIM) project to measure aquifer characteristics but data are only available for a few selected sub-basins across the country (CGWB, 2015).

inclusion of regressor variables due to the unavailability of data about aquifer structure. We expect the influence of sub-surface flow to appear in a regression model as spatially auto-correlated error terms. The variance covariance matrix of the error terms in this case is not a diagonal matrix. We use a "geostastical" model to estimate the variance covariance matrix of the error terms from GWL data. We can then estimate regression coefficients through generalized least squares (GLS) with a known co-variance matrix (Stralberg & Bao, 1999). The predictability of the results from this method depends on the quality of the estimated co-variance matrix (Beale et al., 2010). We describe the use of this method for GWLs data in this chapter.

This chapter is organized as follows; in section 3.2, we introduce some background on geo-statistical modeling and discuss the underlying assumptions and methodological concerns when studying spatial auto-correlation in GWL. In section 3.4 we present the methodology for a rigorous characterization of spatial auto-correlation in GWLs including the process of semivariogram estimation and spatial interpolation via the kriging estimator. We present empirical results using administrative data from the state of Uttar Pradesh (UP) in northern India. We discuss the impact of spatial auto-correlation on estimates of regression coefficients in OLS and describe a method to identify the structure of spatial dependence and use it to estimate regression coefficients through GLS.

3.2 Preliminaries

The science of “geo-statistics” applies the precepts of spatial analysis to the fields of geography, hydrology, ecology and related physical sciences. It is used to model spatially distributed physical quantities such as GWLs, GWL fluctuations, precipitation and pollutant concentration as "stochastic spatial processes" [6] defined over a spatial domain (Matheron, 1963). A stochastic spatial process is a collection of *spatially dependent* random variables indexed by spatial location and defined over either a continuous surface or a set of discrete points (N. Cressie, Cressie, & Sons, 1991; Anselin et al., 2001). Consider a stretch of land represented by a subset of 2D space $D \in R^2$. The GWL G_i

[6] A stochastic spatial process is sometimes also called a "random function" or a "spatial random field."

at every location w_i on this land is a realization of a random variable. At any point in time, all of the GWLs taken together form a continuous groundwater surface which we call the "reality." This surface is considered a single realization of a stochastic spatial process defined over D. The stochastic spatial process is denoted as:

$$\{G(w) : w \in D\} \tag{3.1}$$

where $G(w)$ refers to the collection of indefinitely many random variables and w to an arbitrary location in D (N. Cressie & Moores, 2021).

Spatial autocorrelation is the extent of spatial dependence between the random variables that constitute this process. In simple terms, spatial autocorrelation refers to the functional relationship between GWLs at one location in the domain to those at other locations (Anselin, 2013). There are two ways to characterize spatial autocorrelation; the data-driven approach (N. Cressie et al., 1991) and the model-based approach (Anselin, 2013). The data-driven approach uses point sample data to estimate a semivariogram; a functional representation of the relationship between covariance (or correlation) in GWLs and the "separation" distance between two locations. It does not presuppose any prior relationship between GWLs at different locations. The model-based approach *imposes* a spatial structure through a "spatial weights matrix." The structure of the spatial weights matrix specifies for each location its "neighboring" locations. The elements of the matrix are weights that measure the strength of spatial correlation between GWL at neighboring locations. The neighborhood structure is specified apriori based on knowledge about the particular stochastic spatial process that generates the data.

We follow a data-driven approach in which observed GWL data are seen as samples from a *single realization* of the stochastic spatial process. The problem is we want to pool data over space to be able to estimate the spatial relationships which requires data from multiple realizations of the same process. This problem is handled through the assumption of "spatial stationarity." The validity of geostatistical models is founded on our *decision* to use a spatially stationary stochastic spatial process to represent the observed data. Spatial stationarity is a *property of the model* and not the data. In simple terms, spatial stationarity implies "invariance" of statistical properties over space.

Spatial stationarity implies *constant mean* and *location-independent covariance* over the spatial domain. A stochastic spatial process has a constant mean over a spatial domain if the mean of each random variable in the process is the same. Location-independence of covariance is the idea that the covariance between random variables at any two locations in the stochastic spatial process is independent of their location and dependent only on the "spatial lag" between the two locations. The spatial lag $\bar{h}$ is the displacement, entailing a distance and direction, between two locations[7]. In a spatially stationary stochastic process, given some value of $\bar{h}$, the covariance between all pairs of locations separated by $\bar{h}$ is the same irrespective of their locations.

The use of geostatistical models entails an implicit *decision* to assume spatial stationarity. The decision has to be substantiated by prior knowledge about the natural processes that govern GWL dynamics. We discuss the spatial stationarity decision with respect to GWLs in section 3.2.2. There is generally no way to validate the decision from observed data although exploratory data analysis can be used to rule it out. A clear violation of spatial is the presence of spatial heterogeneity. If exploratory analys reveals the presence of spatial trends or regimes in the data indicating a *systematic* variation in the mean over space, there is evidence to refute spatial stationarity (N. Cressie, 2015).

If the stationarity decision is acceptable, observed data can be pooled over space to estimate spatial structure and relationships. In particular, the stationarity decision enables us to model the spatial auto-correlation using a functional representation called the semivariogram. The semivariogam, usually denoted by $\gamma(\bar{h})$, is a function of variance over spatial lag $\bar{h}$. It gives us a measure of variance between the values of a spatial variable at two locations separated by a spatial lag $\bar{h}$[8]. If a spatial process exhibits positive spatial auto-correlation, we expect that the variance increases with $\bar{h}$ therefore the semivariogram should be an increasing function of $\bar{h}$ and conversely the covariogram to be a decreasing function of $\bar{h}$. We detail the process of semivariogram estimation and modeling in section 3.4.

There are infinitely many random variables corresponding to infinitely many loca-

[7]The term spatial lag is an analogy to the term "lag" in the study of serial auto-correlation in time series data. Though the concepts of spatial and temporal lags are analogous in some ways, they are not identical.

[8]The "covariogram" is the dual of the semivariogram and gives us a measure of covariance between random variables at two locations separated by $\bar{h}$.

tions in the spatial domain D. For convenience, the spatial process in 3.1 is defined through a finite-dimensional joint CDF of n random variables. Equation 3.2 expresses the finite dimensional joint CDF of a stochastic spatial process for some vector of real values $(g_1, ..., g_n)$ (N. Cressie et al., 1991).

$$F_{w_1,...,w_n}(g_1, ..., g_n) = P\{G(w_1) \leq g_1, \ldots, G(w_n) \leq g_n\} \tag{3.2}$$

If we drill n observation wells at locations $w_i,\ i = 1, \ldots, n$ and we make an observation at each well at a fixed point in time then we have a subset of a single realization of the stochastic spatial process. We denote the random variable associated with groundwater level at well w_i as $G(w_i)$ and a single realization of it as $g(w_i)$.

The observed groundwater level data for a cross-sectional dataset is then the set of realizations $g(w_i) :\ i = 1, \ldots, n$. This is all the information we have. We do not know anything about the groundwater levels at locations other than the ones at which we have made observations. Using the available information and the model of a stochastic stochastic process, we accomplish three goals:

1. Spatial Characterization: Characterize the spatial variation including the spatial heterogeneity and spatial autocorrelation (dependence) in groundwater levels across space D.
2. Spatial Prediction: Estimate the values of groundwater levels at unsampled points in D to produce an estimated continuous groundwater surface and compute the confidence intervals of the estimation.
3. Spatial Regression: Estimate the relationship between rainfall and groundwater level while accounting for spatial auto-correlation.

If we allow the random variable at every different location in space to have a different probability distribution and different moments then the model is intractable. We have innumerable parameters to estimate. Further, pooling of data is not possible because the observed data are samples from different populations (Goovaerts, 1997). The spatial stochastic process model is usable only if we make certain assumptions of statistical invariance over space; the property known as spatial stationarity.

3.2.1 Types of Spatial Stationarity

There are varying degrees of how strongly a spatial stochastic process exhibits spatial stationarity. There are broadly three categories of spatial stationarity; strict or strong stationarity, second-order stationarity and intrinsic stationarity.

Strict or Strong Spatial Stationarity

Strict stationarity is the condition that the joint cumulative probability density function (CDF) of the entire family of random variables that constitute the stochastic spatial process is invariant over spatial lag which is expressed in equations 3.3-3.4.

$$F_{w_1,\ldots,w_n} \equiv F_{w_1+h,\ldots,w_n+h} \tag{3.3}$$

$$P\{G(w_1) \leq g_1, \ldots, G(w_n) \leq g_n\} = P\{G(w_1+h) \leq g_1, \ldots, G(w_n+h) \leq g_n\} \tag{3.4}$$

for all values of $\bar{h}$ where $\bar{h}$ is a 2D displacement vector called the spatial lag (Bárdossy, 1997).

Second-order or Weak Stationarity

Strict stationarity is too strong a condition to be plausibly met by natural phenomena. A stochastic spatial process is called "second-order or weakly stationary" when the mean of the family of random variables is constant over the entire region (equation 3.5) of space and the covariance between two random variables separated by a spatial lag $\bar{h}$ depends only on $\bar{h}$ and not on absolute positions (equation 3.6).

$$E[G(w_i)] = \mu \qquad \forall\, w_i \in D \tag{3.5}$$

$$Cov[G(w_i), G(w_i+h)] = C(h) \qquad \forall\, w_i \in D \; and \; h > 0 \tag{3.6}$$

where μ is a constant and $C(h)$ is called the “covariogram” or the "covariance functio" of the process (Bárdossy, 1997).

Intrinsic Stationarity and the Semi-variogram

In practice, a weaker assumption than second order stationarity is used which is called intrinsic stationarity which also requires a spatially invariant mean but not covariance. It requires that the *variance of the differences* between two random variables separated by $\bar{h}$ depend only on $\bar{h}$ and not on absolute positions (equations 3.7-3.8).

$$E[G(w_i)] = \mu \qquad \forall\, w_i \in D \tag{3.7}$$

$$Var[G(w_i + h) - G(w_i)] = 2\gamma(h) \qquad \forall\, w_i \in D \; and \; h > 0 \tag{3.8}$$

where μ is a constant and the functional form $\gamma(h)$ is called the “semivariogram” of the stochastic spatial process (N. Cressie et al., 1991; Bárdossy, 1997).

The semivariogram is function of dissimilarity across space. It is defined as the variance of the differences between pairs of random variables $G(w_i)$ and $G(w_j)$ at spatial displacement $\bar{h}$ from each other (N. A. Cressie, 1993).

$$Var[G(w_i) - G(w_j)] = 2\gamma(w_i - w_j) \qquad \forall\, w_i, w_j \in \mathbb{R}^2 \; s.t \; i \neq j \tag{3.9}$$

where $(w_i - w_j) = h$. The multiplication by 2 is by convention and the variogram refers properly to the quantity $2\gamma(h)$ and $\gamma(h)$ is called the semi-variogram.

The variogram is only defined under the assumption of intrinsic stationarity which is our working assumption for estimation. An "experimental variogram" is first estimated from the available data and then fitted with a theoretical model.

3.2.2 Spatial Stationary Decision for Groundwater Levels

Many models can be used to represent the same reality and the choice depends on the goal at hand (Goovaerts, 1997). From the perspective of model selection, the "stationarity decision" is a decision by the analyst or researcher to use spatially stationary stochastic process as a mental model for the reality. In this sense, the stationarity decision points to a property of the analyst's mental model. It is the property of *invariance* in a set of statistical properties of the chosen stochastic process model over a specified spatial domain as expressed in equations 3.3 - 3.8.

Our interest lies in identifying and characterizing the structure of spatial auto-correlation in groundwater levels. If we *decide* to employ a stationary stochastic spatial process to do this, we must first *argue* using prior knowledge that our decision is a reasonable one. From the perspective of the represented reality which in our case is groundwater dynamics, the stationarity decision *implies* but does not validate the assumption that groundwater levels fluctuate within an "unchanging envelope of variability (Milly et al., 2008)." This is to say that we consider the groundwater dynamics in a region to be homogeneous in the sense that we find it reasonable to use a *single* process to model their spatial variations.

References to the concept of "stationarity" thus refer both to an *explicit* decision to use a stationary spatial process model and an *implicit* assumption of homogeneous variability in the modeled reality. The decision, in general, cannot be verified or refuted using observed data (Goovaerts, 1997) since the data constitutes only a single realization of the stochastic process. There is no standard method or statistical test to validate the stationarity decision and arguments must follow the principles of logical reasoning (Goovaerts, 1997).

Arguments for the validity of the stationarity decision in the spatial modeling of groundwater level variations can come from two sources. First from prior knowledge about the hydrologic cycle as it operates in a given region and second from exploratory data analysis.

Exploratory analysis can reveal certain patterns of variation in the data that invalidate the stationarity decision. The first and perhaps the easiest to detect are "spatial

trends." A salient feature of a spatially stationarity stochastic process is that variation is restricted to some *localized* spatial "range" that is small compared to the extent of the entire domain under consideration. Trend is a term carried over from time-series analysis and in the spatial context it signifies a continuous, *global* variation across the domain. Trend detection and removal is an important part of geo-statistical modeling. If a spatial trend is detected in the data, a variant of kriging called "universal kriging" is used in which spatial modeling and prediction proceed on the de-trended or de-meaned data and the trend component is added back in to produce the finally predicted values.

Another indication that stationarity is not valid is if exploratory data analysis points to the existence of more than one distinct groundwater level regime with significantly different statistical parameters. The way to handle this is to sub-divide the region into sub-regions within which statistical parameters remain the same (Goovaerts, 1997).

Intrinsic stationarity is usually considered sufficient geo-statistical modeling because it imposes the weakest restrictions on the properties of the stochastic process (see equations 3.7-3.8). It entails a constant mean and the sole dependence of the variogram *only* on spatial lags and not absolute positions.

In general, soils and hydro-geological characteristics are reported as being uniform over the entire Gangetic basin characterized by unconfined alluvial aquifers. Bhanja has published a map of aquifer types for India which shows that most of the Indo-Gangetic Basin and practically all of UP fall within the same category of unconsolidated sedimentary aquifers (Bhanja et al., 2016). Bonsor *et. al* have integrated findings and data from over 500 studies and synthesized 7 distinct "groundwater typologies" that constitute different groundwater regimes across the Indo-Gangetic Basin (Bonsor et al., 2017) (see Figure 3.2.1). The entire region receives most of its rainfall from the south-west Monsoon during the months of May, June, July and August although the amount of rainfall shows an increasing trend from west to east. From the perspective of climatic and hydro-geological processes, the region can be said to possess a degree of uniformity which supports the stationarity decision.

Human factors, on the other hand, vary considerably across the state of basin. Singh *et. al* have published a high resolution map of cropping patterns for Uttar Pradesh that shows clustered regimes of cropping distributed roughly according to established agro-

climatic zonal divisions of the state (see Figure 3.2.2) (Singh, Kudrat, Jain, & Pandey, 2011). Human activity is considered by some experts to challenge the validity the stationarity decision due to unprecedented impacts of anthropogenic processes on climate and hydrologic systems that introduce extreme variations in the means and extremes of precipitation (Milly et al., 2008). Others have argued that non-stationarity does exist but the complexity and scale of hydroclimatic systems makes it impossible to conclusively establish its existence via statistical methods and the scientific community can benefit from the stationarity decision in hydrology if due diligence is done as to its consequences for estimation and inference (Lins & Cohn, 2011).

Our arguments for the stationarity decision for groundwater level dynamics are supported by the relative hydro-geological uniformity of the region and the reported positive spatial auto-correlation in groundwater levels by other researchers but challenged by the presence of diverse cropping regimes and extremes in groundwater levels across UP. As a middle ground, we propose to cluster the state by agro-climatic zone which is a spatial grouping that accounts both for the climatic and cropping processes active in sub-regions. As a refinement to this basic grouping, we may further divide regions within each zone according to what we term the "hydrologic schedule." The hydrologic schedule is a schematic diagram of the processes of extraction and recharge in a given cropping year for a particular sub-region. While recharge occurs approximately during the same time for the whole state, extraction patters vary by cropping system and these induce different hydrologic schedules which can be used group similar sub-regions into supposedly stationary spatial domains (see Figure 3.2.1).

3.3 Review of Literature

The idea that “everything is related to everything else but nearby things are more related than distant ones” is credited to Tobler who applied it to study urban growth using a simulation. This adage is known as Tobler’s Law or sometimes as the First Law of Geography (Tobler, 1970). Tobler proposed a second law that states that “things that are external to the area of interest affect things that are inside (Tobler, 1999).” Tobler’s laws are applied to problems in economics through formalisms encoded in the

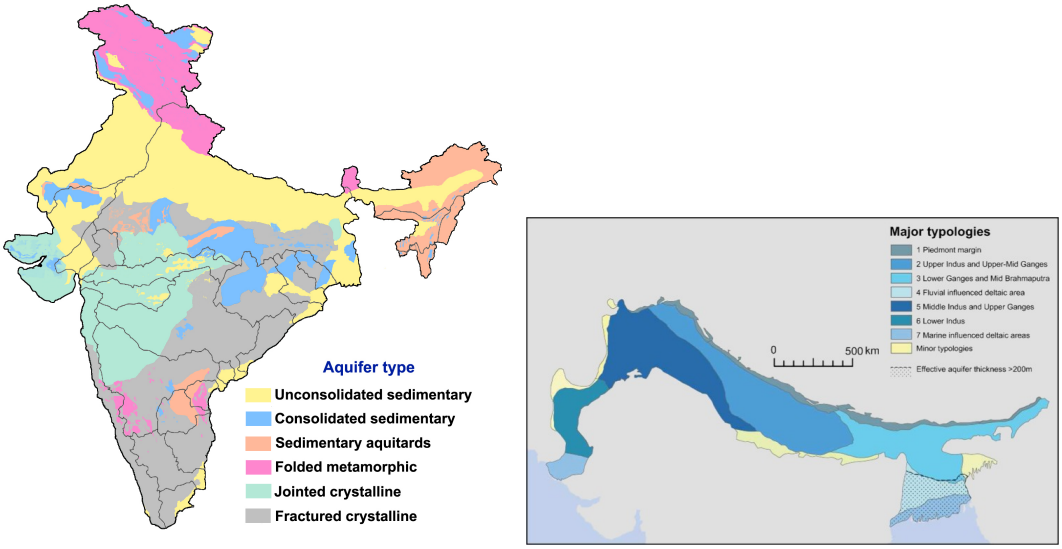
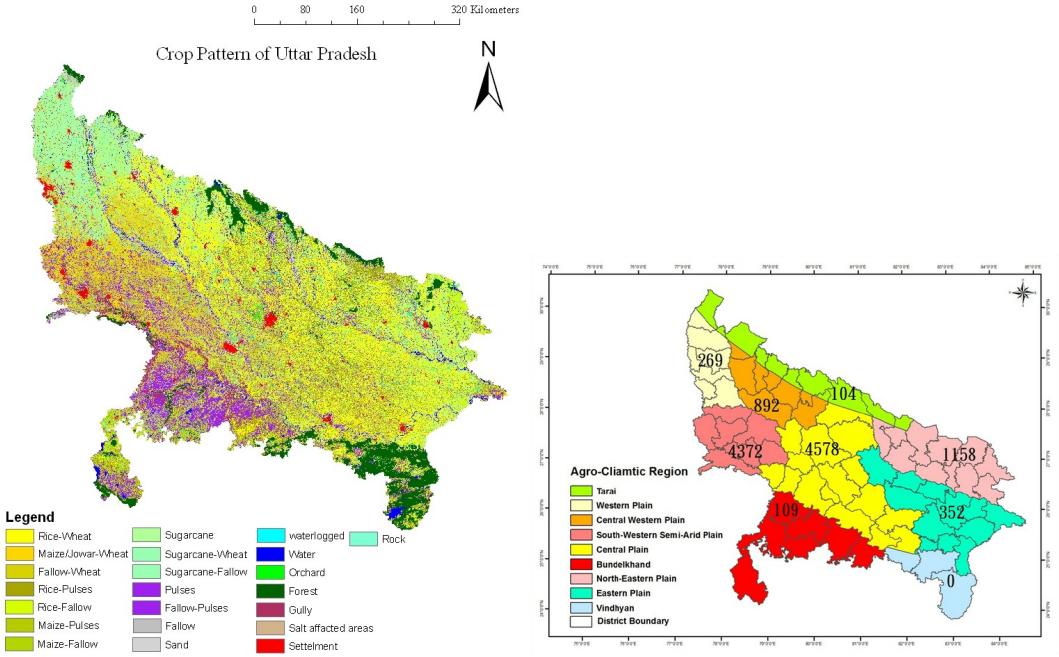


Figure 3.2.1: Left: Aquifer Types in India (Bhanja et al., 2016). Right: Groundwater Typologies of the Indo-Gangetic Basin (Bonsor et al., 2017)

science known as "spatial statistics." The term spatial statistics refers to the application of statistical concepts and methods to data that have a spatial location attached to them, and in which this locational element is used as an important and necessary part of the analysis (Unwin, 2009).

Geo-statistics is a type of spatial science developed by Matheron to characterize the spatial distribution of geographical variables (Matheron, 1963). Spatial econometrics deal with specification, testing, and estimation of empirical models in situations where spatial effects are present (Anselin et al., 2001; Lauridsen, 2012). Historically dating back to the 1970s, interest in spatial econometric models has grown on account of new theoretical frameworks of "interacting agents" in economics that model strategic interaction, social norms, neighborhood effects, copy-catting, spatial externalities, spillovers and other peer group effects that address questions regarding how individual interactions can lead to emergent collective behavior and aggregate patterns (Anselin et al., 2001).

Space is incorporated into econometric models either through model-driven or data-driven approaches. Model-driven approaches are based on prior knowledge and *begin* with a model or theory about the spatial structure of the data (Anselin, 1988). An example is the *spatial autoregressive* (SAR) model in which a weighted spatially lagged variable is included as a regressor to account for spatial auto-correlation and the analyst



Example of a hydrologic schedule

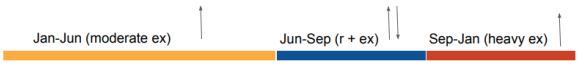


Figure 3.2.2: Looking for Spatial Stationarity: Cropping Patters (Top-Left) and Agro-Climatic Zones (Top-Right), Uttar Pradesh. An example of a hydrologic schedule for a hypothetical sub-region (Bottom). Sources: Singh *et. al* (Singh et al., 2011), Jha *et. al* (Jha & McKinley, 2014)

must specify before hand the weights matrix that determine how neighboring observations are related (Anselin et al., 2001). By contrast, data-driven approaches begin by identifying or estimating the spatial structure from observed data (N. Cressie et al., 1991).

In the context of linear regression, spatial auto-correlation may occur in the dependent variable, one or more of the regressors, the unobserved error terms or some combination of these. The most likely situation that will be encountered in ecological economics is of spatially auto-correlated error terms (Beale et al., 2010). If all of the spatial-autocorrelation in the dependent variable is explained by the spatial auto-correlation in the regressors, this is usually does not pose a problem for estimation (Beale et al., 2010). This is generally not the case in practice when working with spatial data and the presence of unexplained *residual* spatial auto-correlation in the error terms is very likely (Beale et al., 2010). Spatially auto-correlated error terms violate the classical assumption of independence and potentially vitiate statistical inference (Legendre, 1993; F. Dormann et al., 2007).

Generally, the true structure of spatial auto-correlation is unknown and its effects on the estimates of regression coefficients cannot be verified using observed data and it is common practice to study such effects using simulation (Beale et al., 2010). Beale *et. al* have demonstrated that Type 1 error rates increase in the case of non-spatial regression methods like ordinary least squares (OLS) with increases in autocorrelation in the errors but not when it is explicitly accounted for via generalized least squares (GLS) (Beale et al., 2010). If spatial auto-correlation is ignored then the estimates of regression coefficients have greater standard errors (N. A. Cressie, 1993). Beale *et. al* suggest that this reduced precision could be the reason behind reports in the ecology literature of point estimates from GLS being smaller in magnitude than those obtained with OLS. In this work, we study the relationship between rainfall and groundwater level using a hybrid approach that combines real world data with simulated error terms.

Variations in groundwater recharge and extraction at a particular location are a result of complex spatial interactions of human entities and natural processes occurring locally and remotely across space. A rigorous understanding of spatial interactions and effects in groundwater level dynamics and factors affecting those dynamics is important for

groundwater management and policy. Spatial econometric and geostatistical models have been applied to the studying the impact of spatio-temporal land-use change on groundwater depletion (Moukana & Koike, 2008), modeling spatial interdependence among water users (Ekpe, 2021), estimation of water demand from depleting aquifers (Jamali Jaghdani, 2012), estimation of spatial pumping externalities (Farsi aliabadi, Daneshvar Kakhki, Sabouhi, Dourandish, & Amadeh, 2020), spatial interpolation of groundwater levels (Delhomme, 1978) and imputing missing values in groundwater level data (Chung et al., 2019).

3.4 Materials and Methods

Exploratory analysis of spatial data often begins with the computation of the Moran's I coefficient of spatial auto-correlation (Moran, 1950; Bivand, Pebesma, Gomez-Rubio, & Pebesma, 2013) also known as the Moran's *i* statistic. The statistic tests for significant correlation between "adjacent" observations where adjancency is pre-specified through a spatial weights matrix that assigns weights to each neighbor. Let $w_{i,j}$ be the "strength" of the correlation between well i and well j. Using all pairs of wells we encode the spatial relationship into a spatial weights matrix $[W]_{N\times N}$ where $W[i,j] = w_{i,j}$. The statistic is computed by regressing groundwater level $G_{N\times 1}$ against a spatially lagged variable $WG_{N\times 1}$. The Moran's I statistic is the slope of the regression line.

The statistic can be tested by Monte Carlo simulation under a null hypothesis of spatial randomness. The Moran's I statistic is computed by repeatedly assigning random weights for n iterations. This produces a randomized distribution of Moran's I statistics which is compared to the observed Moran's I and a p-value is computed to test the hypothesis of spatial randomness (Bivand et al., 2013).

While the Moran's I gives us a "global" measure of spatial auto-correlation in the data, it requires a prior notion of adjacency, conveys nothing about the spatial range or anisotropy in the structure of spatial auto-correlation. We thus resort to the more powerful instrument for identifying spatial structure called the semivariogram or simply the variogram.

3.4.1 Variogram Estimation

If stationarity is acceptable on logical grounds then the variogram is well defined and can be estimated from the data and then fitted with a theoretical model that represents the spatial structure present in the data. We detail this process in the following sections.

A variogram estimated from data is called an *experimental variogram* and denoted $2\hat{\gamma}(h)$. A number of variogram estimators exist but the most commonly used is the method-of-moments estimator also known as the "classical" estimator (N. Cressie et al., 1991). Under the intrinsic stationarity assumption, the experimental variogram $2\hat{\gamma}(h)$, of a stochastic spatial process can be estimated by equation 3.10.

$$2\hat{\gamma}(h) = \sum_{|h|} \frac{(g_i(w_i) - g_j(w_j))^2}{n(|h|)} \tag{3.10}$$

where $|h|$ is the set of all pairs of wells w_i, w_j that are separated by a displacement h and $n(|h|)$ is the number of wells in the set $|h|$. In practice, we generally do not find wells that are separated by exactly a displacement h so we allow a tolerance of d and consider wells that lie at a distance within the interval $[h - d/2, h + d/2]$. Figure 3.4.1 illustrates the computation of an omni-directional experimental variogram. We consider pairs of wells moving outward in all directions relative to the well w_4. The experimental variogram is then computed at each h_i using the expression in 3.10. This process must be repeated for each well in the data without double counting of point pairs.

It is also possible to compute directional or anisotropic experimental variograms where pairs of wells are considered only in a certain direction $\bar{h}$. This is shown in Figure 3.4.1 (bottom panel). For directional variograms, we have to introduce an angular tolerance α because in practice we do not find wells that lie exactly along a line. We consider wells lying within the cones of angle α.

3.4.2 Variogram Model Fitting

The experimental variogram gives us an estimate of variance at a set of discrete values h_i which cannot be directly used for spatial prediction (see section 3.4.3). A theoretical

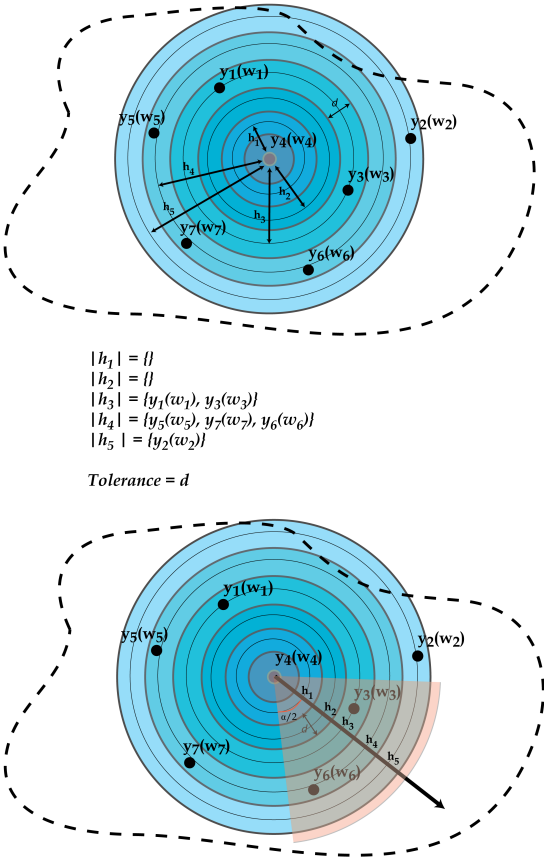


Figure 3.4.1: Top: Computation of an omnidirectional or isotropic experimental variogram. Relative to the well w_4, we move outward at distances h_1, h_2, ... h_5 and allow a tolerance d. Bottom: Computation of a directional or anisotropic experimental variogram. Only the region shaded in red is considered for computation but the procedure is the same as that for an omnidirectional variogram. Recall that this procedure must be repeated for each well.

model that most closely resembles the spatial dependence structure estimated from the data is fitted to the estimated variogram (N. Cressie et al., 1991).

Model fitting via maximum likelihood depends on the assumption that the original data Y are independent and multivariate Gaussian (N. Cressie et al., 1991). The most common method of variogram model fitting is to fit a chosen parametric family of variograms. Typical functional forms used for fitting are the linear, exponential and spherical model variograms. Once a suitable fit is obtained, the fitted variogram is called the "theoretical variogram."

3.4.3 Spatial Prediction of Groundwater Levels

The goal of spatial prediction is to estimate the groundwater level at an unsampled location w_0 using available data sampled at locations $w_i, i = 1, ..., n$. The technique known as "ordinary kriging" estimates the value at w_0 as a weighted linear sum of the values at sampled locations (see equation 3.11). The derivation of kriging equations and the estimation variance in this section is due to Shortridge (Shortridge, n.d.).

$$\hat{y}_0(w_0) = \hat{y}_0 = \sum_{i=1}^{n} \lambda_i g_i(w_i) = \begin{bmatrix} \lambda_1 & \dots & \lambda_n \end{bmatrix} \begin{bmatrix} g_1(w_1) \\ \vdots \\ g_n(w_n) \end{bmatrix} \tag{3.11}$$

where $g_i(w_i)$ is the i^{th} sampled value and λ_i the weight associated with it.
The kriging estimator incorporates the covariance structure among the data samples to compute the weights for predicting the value $Y_0(w_0)$. The weight are computed based on two covariances:

1. The covariance *within* the sampled data points:
   $C_{i,j} = C(w_i, w_j) = Cov(g_i(w_i), g_j(w_j)) \ \forall \ i, j$ which are expressed as a covariance matrix:

$$C_1 = \begin{bmatrix} C_{1,1} & C_{1,2} & \dots & C_{1,n} \\ C_{2,1} & C_{2,2} & \dots & C_{2,1} \\ \vdots & \vdots & \ddots & \vdots \\ C_{n,1} & C_{n,2} & \dots & C_{n,n} \end{bmatrix}$$

2. The covariance *between* the sampled data points and the unsampled location: $C_{i,0} = Cov(g_i(w_i), Y_0(w_0)) \; \forall \; i$ which are expressed as a covariance vector:

$$c_0 = \begin{bmatrix} C_{1,0} \\ C_{2,0} \\ \vdots \\ C_{n,0} \end{bmatrix}$$

These covariance terms are estimated using the modeled variogram. Kriging weights λ_i are computed under the following constraints:

1. Unbiasedness
   $E[\hat{G}_0(w_0)] = E[G_0(w_0)]$
   This will be satisfied if $\sum \lambda_i = 1$ and the mean is stationary i.e $E[G(w_i)] = \mu \; \forall \; i$, where μ is a constant.

2. Minimum Variance
   The kriging estimator is the optimal estimator in the sense that it minimizes the *prediction variance*, σ_ϵ^2, which is the variance of the difference between the theoretical "true" value $g_0(w_0)$ and the estimated value $\hat{g}_0(w_0)$. The prediction variance is also known as the mean squared prediction error (MSPE).

   $$\sigma_\epsilon^2 = Var[g_0(w_0) - \hat{g}_0(w_0)] \tag{3.12}$$

   Lagrangian minimization yield the kriging equations for the weights. I omit the derivation here for brevity but refer the reader to the text by Shortridge (Shortridge, n.d.). The kriging equations are given below in matrix form:

   $$W = \begin{bmatrix} \lambda_1 \\ \lambda_2 \\ \vdots \\ \lambda_n \end{bmatrix} = \begin{bmatrix} C_1 & 1 \\ 1' & 0 \end{bmatrix}^{-1} \begin{bmatrix} c_0 \\ 1 \end{bmatrix} = C^{-1}D \tag{3.13}$$

   where C_1 is the covariance matrix containing the covariances within the data locations and c_0 is the covariance vector containing the covariances between the data and the unknown locations, W is the vector of kriging weights.

The kriging variance is given by:

$$\sigma_\epsilon^2 = \sigma^2 - W'D \tag{3.14}$$

3.4.4 Geo-statistical Modeling Procedure

We now summarize the entire geo-statistical modeling procedure. It typically consists of the following steps though some variations of it maybe followed depending on the application.

1. **Outlier Detection:** In the spatial context, points that represent large differences from other points nearby are classified outliers.
2. **Anisotropy Detection:** Anisotropy shows up in the data as a systematic trend where values vary by direction. If the data appears to be anisotropic then a directional variogram has to be used instead of an omni-directional one.
3. **Variogram Estimation:** As outlined in section 3.4.1.
4. **Variogram Model Fitting:** As outlined in section 3.4.2.
5. **Kriging:** Once a suitable theoretical variogram has been obtained, it can be used to perform spatial prediction through kriging (section 3.4.3). Kriging can then be used to prepare spatially interpolated maps of the groundwater level.
6. **Model Cross-Validation:** Using the kriging estimator, it is possible to cross-validate the chosen theoretical variogram model. Cross-validation removes a subset of the observed data and then estimates it using the remaining data and provides a measure of accuracy by comparing the results with original values.

$$Var[\hat{\beta}_{ols}] = \sigma^2(X'X)^{-1}(X'\Omega X)(X'X)^{-1} \tag{3.15}$$

Where Ω is the variance covariance matrix of the autocorrelated error terms. We need to estimate the matrix Ω to be able to compute the variance of the OLS estimator and also to perform GLS estimation. It turns out we can use the spatial structure estimated by the theoretical variogram to obtain the variance covariance matrix of the spatially autocorrelated errors.

3.4.5 Identifying Spatial Structure through the Semivariogram

The structure of spatial auto-correlation in the groundwater level G and as a consequence in the errors u is expressed by the off-diagonal elements of the variance covariance matrix of the error terms $\Omega = E[u'u]$. In general, structure of this matrix is unknown but in the special case of spatial auto-correlation, the structure of spatial

dependence can be estimated from the observed data through the experimental semivariogram and then the semivariogram can be used to obtain the estimated variance covariance matrix $\hat{\Omega}$. This is the approach that is generally known as feasible generalized least squares (fGLS) which can be applied here to estimate regression coefficients and standard errors.

We have already explained semivariogram estimation in the section 3.4.1 and fitting a theoretical model to the experimental semivariogram in section 3.4.2. At the end of the process of variogram estimation and modeling, we have a closed functional form (the theoretical semivariogram) $\gamma(h)$ that expresses the structure of spatial dependence as obtained from the observed data. For any two observations g_i and g_j separated by a spatial lag (and direction) h, the covariance between the corresponding error terms is given by equation 3.16 (N. Cressie et al., 1991).

$$Cov[u_i, u_j] = C(h) = C(0) - \gamma(h) \tag{3.16}$$

We can use the theoretical semivariogram $\gamma(h)$ to obtain estimates of the off-diagonal terms of the variance covariance matrix of the error terms as in equation 3.17.

$$\hat{\Omega}_{i,j} = C(0) - \gamma(h) \tag{3.17}$$

where $dist(u_i, u_j) = \bar{h}$ is the displacement between the spatial locations of the observations g_i and g_j. The coefficient β in the model given by 2.25 can then be estimated using the GLS form in 3.18.

$$\hat{\beta}_{GLS} = (X'\hat{\Omega}^{-1}X)^{-1}X'\hat{\Omega}^{-1}G \tag{3.18}$$

Equation 3.18 makes it clear that the estimated matrix $\hat{\Omega}$ must be invertible.

3.5 Simulation Framework

We employ Monte Carlo simulation to study the impact of spatial auto-correlated errors on estimates of linear regression coefficients. The goals of the simulation are:

1. To assess the impact of the presence of spatially auto-correlated error terms on OLS estimators relative to normally distributed i.i.d errors
2. To assess the impact of the range of spatially auto-correlated error terms on OLS estimators
3. To assess the relative efficiency and unbiasedness of variogram-based fGLS estimation compared to standard OLS estimation.

Consider a situation where we have N real-world observations each of total monsoon rainfall X (mm) and post-monsoon groundwater level G (m). We expect that groundwater levels are spatially auto-correlated[9]. We generate two sets of error terms; first a set that are i.i.d and normally distributed and second that are spatially auto-correlated with a know spatial structure. We then combine these simulated error terms with real-world rainfall data using the true model expressed in equation 2.25 to get simulated two corresponding sets of GWLs. We then use the two sets of data to compare the behaviour of OLS estimators in the absence and presence of spatially auto-correlated errors. Note that the spatial structure of GWLs, the dependent variable, is the same as the errors. Since errors are linearly related to groundwater level through the true model expressed by 2.25, under classical assumptions, the unknown variance-covariance matrix of the errors Ω is equal to that of the groundwater levels.

We also evaluate the impact of the range of spatial auto-correlation on OLS estimation bias and efficiency relative GLS. Recall that the range of spatial auto-correlation is the distance beyond which covariance between two random variables in space is zero. We study the relative efficiency of OLS and GLS estimators with respect to variation in range. This is important for groundwater studies and policy since the range differs from one aquifer to another. We let the range vary from approximately 1 km to 50 km. Recall that the estimated range for the study region in this chapter based on the experimental variogram estimated using real GWLs data was 20 km. An interval from 1 km to 56

[9]Note that we also expect there to be spatial auto-correlation in rainfall but for this study, we concern ourselves with GWL as the dependent variable.

km covers the range of spatial auto-correlation encountered in all the studies we have reviewed in this chapter.

Finally, we also study the efficacy of fGLS estimators to address the issue of spatial auto-correlation. We simulate spatially auto-correlated errors using a known or true variogram model with pre-specified model. We set the parameters of the true variogram, sill and range, to values obtained from our experimental variogram model. We then simulate data using the specified spatial structure and generate our GWLs data. We then estimate the variogram again from the simulated data and use the fitted variogram model to obtain the variance-covariance matrix of the error terms. This matrix is used to get fGLS estimates. We then compare the fGLS estimates with the OLS estimates to draw our conclusions. The Monte Carlo simulation procedure is outlined below.

1. Let N_{iter} be the number of Monte Carlo iterations to run and n be the number of data samples in the real-world GWL and rainfall data.
2. **Baseline Estimates:** Regress real-world data for G on real-world data for X using OLS to obtain the baseline estimate of the slope coefficient $\hat{\beta}_{ols}^{rw}$. Note that the estimated variance of this estimator is biased therefore we do not consider it.
3. **True Model**: Posit a true model as expressed by 2.25 and set $\beta = \hat{\beta}_{ols}^{rw}$.
4. **True Spatial Structure**: Generate a true variogram model using the spherical spatial structure and sill (global variance) equal to the sample variance of the real-world GWLs data $\hat{\sigma}^2$.
5. **START** Monte Carlo Iteration (Repeat Steps 5-12 N_{Iter} times)
6. **Simulate Independent Errors ($u \sim N(0, \hat{\sigma}^2 I)$):** Simulate n normally distributed i.i.d errors. Use real-world rainfall data and add the independent normally distributed error vector with zero mean and variance $\hat{\sigma}^2$ to calculate the value of the hypothetical groundwater level using the posited true model. Denote this GWL data G_{iid}.
7. **Simulate Spatially Auto-Correlated Errors** ($u \sim N(0, \Omega_{N\times N})$): Simulate n spatially auto-correlated error terms with a spatial structure represented by the true spherical variogram model. Denote the variance-covariance matrix of these errors as $\Omega_{N\times N}$. Add these errors to the real-world rainfall data to calculate the value of the hypothetical groundwater level using the posited true variogram model. Denote this GWL data as $G_{sp-auto}$
8. **OLS with iid errors:** Estimate the model parameters using OLS by regressing the G_{iid} on the real-world rainfall data. Record the value of the slope estimator and its variance. Let these be $\hat{\beta}_{ols}^{iid}$ and Var$[\hat{\beta}_{ols}^{iid}]$.
9. **Variogram Estimation:** Estimate and fit a variogram model $\hat{\gamma}^1$ to the spatially auto-correlated groundwater level data $G_{sp-auto}$.

10. **Obtain the Variance-Covariance Matrix of Error Terms:** Use the fitted variogram model from step 8 to construct $\hat{\Omega}_{N\times N}$ as specified in section 3.4.5.

11. **OLS with spatially auto-correlated erros:** Estimate the model parameters using OLS and record the value of the slope estimator and its variance. Let these be $\hat{\beta}_{ols}^{sp-auto}$, Var$[\hat{\beta}_{ols}^{sp-auto}]$.

12. **GLS with spatially auto-correlated erros:** Estimate the model parameters using GLS where the covariance matrix is set to $\hat{\Omega}_{N\times N}$ as obtained in step 10. Record the value of the slope estimator and its variance. Let these be $\hat{\beta}_{gls}^{sp-auto}$, Var$[\hat{\beta}_{gls}^{sp-auto}]$.

13. **END** Monte Carlo Iteration

14. Compute the means and variances of all estimators and compare.

3.6 Results

We use administrative well-level data containing groundwater level (GWL) observations made during the post-monsoon (Aug-Sep) period for Uttar Pradesh in the year 2009. Each row in the data corresponds to a well location and a year of observation. We also use gridded (9 km by 9 km) rainfall (RFL) data from the ERA5 Land climate database (Muñoz Sabater, 2019). We remove observations where GWL observations are missing and an outlier where groundwater level was 360 m. Observations are made at two types of wells, shallow wells that reach up to an average depth of 13.5(11.4) m and deep wells that reach up to 31.5(13.9) m. We retain only deep wells. For reasons explained in the next section, we retain only wells belonging the western half of the state which is composed of four agro-climatic zones; the Tarai, the Western Plains, the Mid-Western Plains and South-Western Semi-Arid zones (see Figure 3.2.2 for agro-climatic zones). Table 3.6.1 provides summary statistics and Figure 3.6.1 shows spatial and frequency distributions.

3.6.1 Spatial Stationarity Decision

There is significant spatial non-uniformity in groundwater level across the state of Uttar Pradesh. There is an increasing trend with increasing distance from the foothills of the Himalayas in the North-Eastern part of the state. Monitoring is also non-uniform

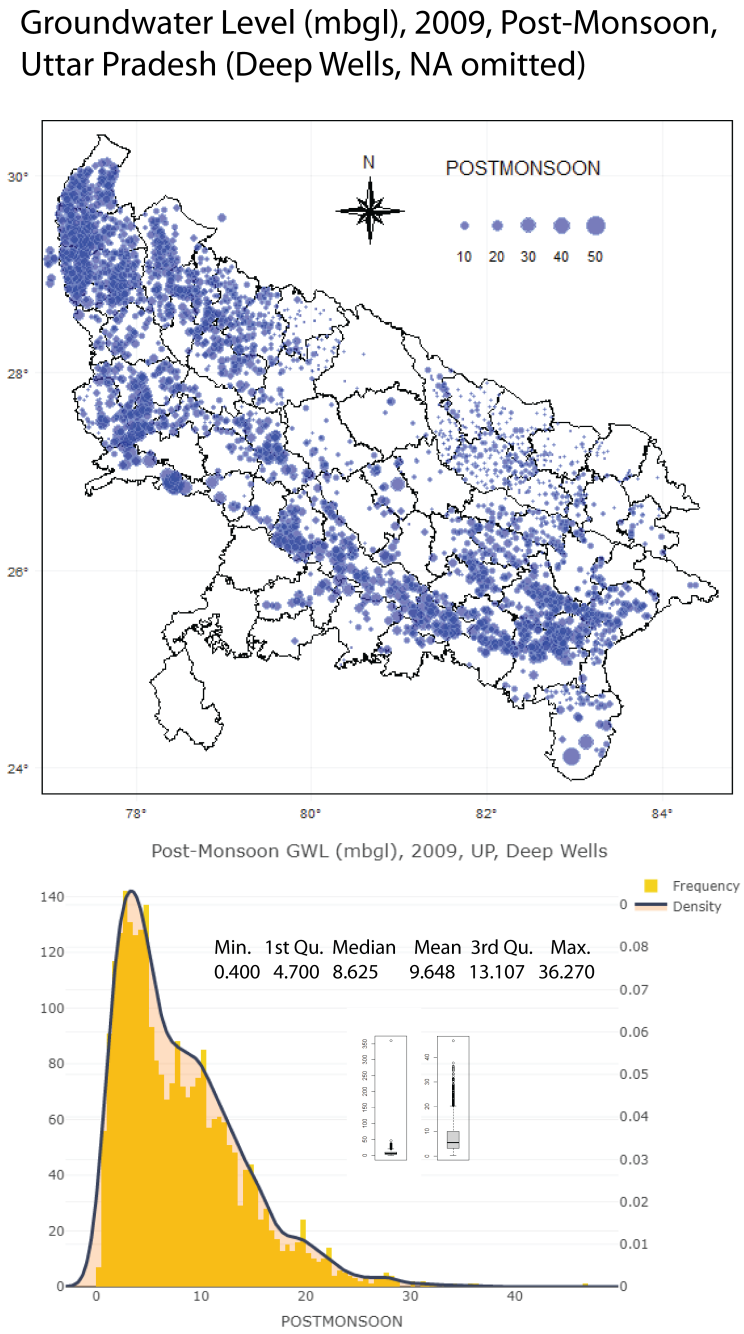


Figure 3.6.1: Spatial (top) and Frequency (bottom) distributions of groundwater level, post-monsoon, Uttar Pradesh, 2009, Deep Wells. Missing values and outliers removed.

with more wells in the Western half than in the Eastern half of the state. While aquifer typology is uniform across Uttar Pradesh, there are significant agro-climatic variations. The Eastern parts receive significantly more rainfall and cropping patterns and by consequence groundwater extraction patterns differ in each zone (Figure 3.2.2). As an exploratory step, to determine the degree of spatial auto-correlation we first computed Moran's I statistic for the entire state, for each agro-climatic zone and for varying sizes of spatial neighborhoods ranging from 1 km to 50 km (see Figure 3.6.2).

As the top panel of Figure 3.6.2 shows, spatial auto-correlation increases with neighborhood size for the whole state up to a range of approximately 9 km and reduces after that. Spatial auto-correlation in the four selected zones (Tarai, Western, Mid-Western and South-Wester Semi Arid) exhibits a similar pattenr attaining a maximum at nearly the same range of approximately 9 km. Spatial auto-correlation is not significant in other zones. Due to the similarity in the Moran's I statistics, lower heterogeneity in groundwater levels and higher availability of data we selected the agro-climatic zones lying in the four agro-climatic zones lying in the Western half of the state and made the decision to model the observed data in these parts as realizations of a spatially stationary stochastic process.

3.6.2 Variogram Modeling and Kriging

1. Step 1: Detecting outliers
   An abnormally large value of 360 m was removed from the data on account of the fact that this value exceeds the maximum reported aquifer thickness of the top aquifer layer in the region (CGWB, n.d.). Figure 3.6.1 shows (inset in bottom panel) box-plots of groundwater levels with all the data and with the outlier removed.

2. Step 2: Variogram Estimation
   Spatial Detrending: We assume spatial stationarity within the four selected agro-climatic zones but even within these zones, there is a noticeable spatial trend in the North-East to South-West (45 degree) direction (see Figure 3.6.1). We control for this trend while estimating the variogram. We first compute a conditional mean groundwater level given the latitude and longitude of well locations and subtract this mean from the groundwater level values. The variogram is computed on this "de-trended" data.

   Width and Cutoff: Recall our aim is to compute an estimate of variance at varying spatial lags. The "width" is the size of the increment in spatial lag and "cutoff" is the maximum lag within which variance estimations are made. These

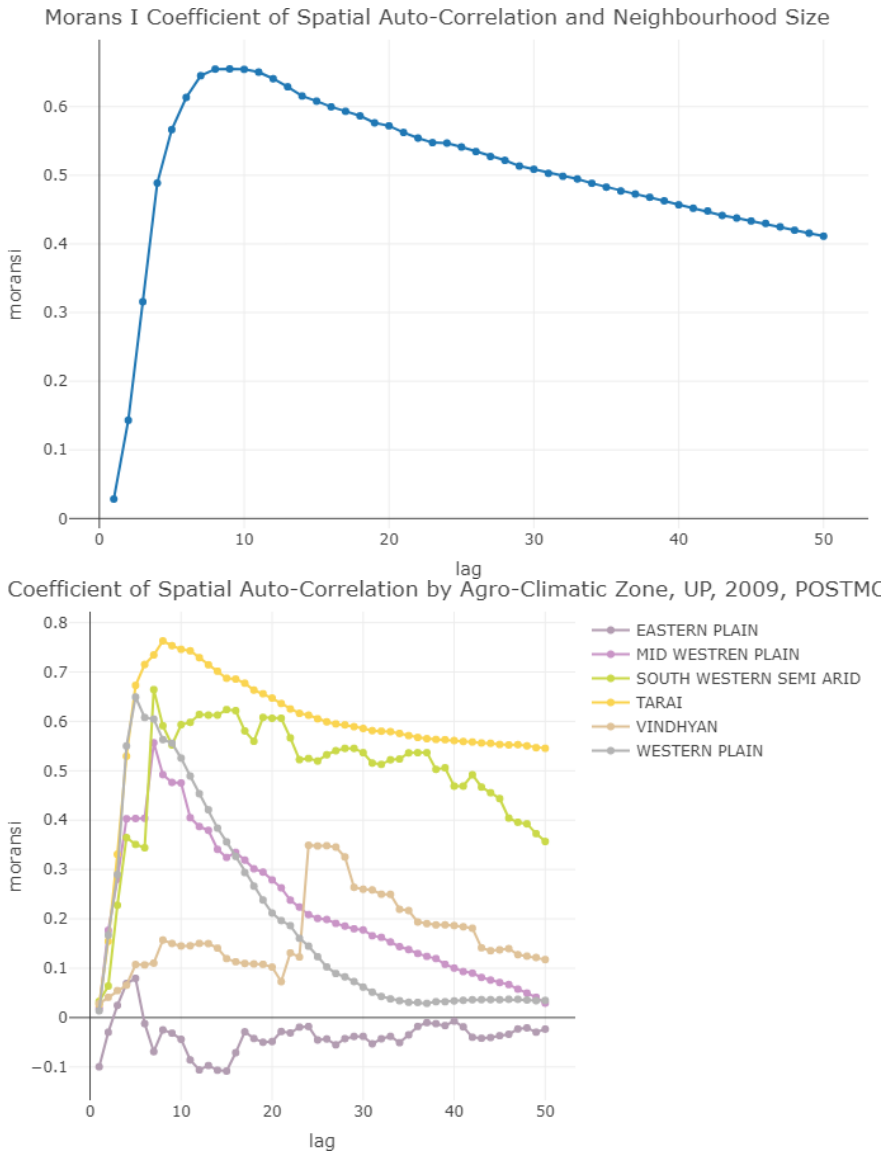


Figure 3.6.2: Moran's I statistic for varying sizes of spatial neighborhoods ranging from 1 km to 50 km for Uttar Pradesh (top) and separately for each agro-climatic zone (bottom). For some agro-climatic zones, there was not enough data to report Moran's I.

factors must be carefully specified because spatial variation is sensitive to scale and resolution. A value for width that is too small will result in over-fitting and the capturing of noise. A value too large will cause us to "smooth over" interesting spatial variation between two increments. We use a width of 5 km which gives us 6 increments between two wells on average. Wells are separated on average by a distance of 30 km but we want to capture spatial variation between them. We use a cutoff distance of 300 km meaning that we assume spatial covariance to be zero beyond this distance.

Anisotropy: Variograms can be isotropic (omni-directional) or anisotropic (directional). The decision of which one to use depends on whether there is reason to believe that direction plays an important role in determining the spatial structure of the data. Figure 3.6.3 shows experimental and model semivariograms for the whole state and for selected agro-climatic zones.

The spatial lag in kilometers is on the x-axis and the expected (semi) variance in meters squared on the y-axis (Pebesma, 2004). The semivariance is in meters squared since the variable of interest is groundwater level which is measured in meters. The results reveal a spatial structure where variance is smaller at small distances and increases with distance till a maximum is reached and flattens out thereafter. This is suggestive of positive spatial auto-correlation in the data which is bounded in the sense that at sufficiently large distances, the variance tends towards a constant which is called the "sill" of the variogram.
In the top panel, we have the directional semivariograms in the four cardinal directions 0, 45, 90 and 135 degrees. The variance in the zero degree direction does not seem to converge to some maximum value. This indicates the presence of a residual trend even in the detrended data. The other three directions produce similar semivariograms. In Figure 3.6.1 we may notice that the 135 degree direction is the direction of "maximum continuity." If we draw parallel lines in this direction, we are likely to encounter very similar values of groundwater level. In practice, this direction is preferred for modeling directional semivariograms as it is likely to be least affected by spatial trend. The bottom panel of Figure 3.6.3 shows 135 degree semivariograms for the whole state (left) and for the selected agro-climatic zones (right). The range and variance in the zonal variograms is less than that for the whole state indicating a smaller scale of spatial auto-correlation in these zones as compared to the entire region. We use the zonal 135 degree semivariogram fittend by a spherical model with range 20 km for the rest of the analysis.

Data Availability: To increase confidence in the estimated variogram, we must have a sufficient number of point pairs at each value of spatial displacement h. In general it is suggested that at least 30 pairs are required to get a useful estimate (Bárdossy, 1997). We have more than 30 points for all lags within the range of 20 km.

3. Step 4: Variogram Model Fitting
We use a spherical model to achieve the best fit. The best fit model had a nugget effect of 2.4 meters squared, a sill of 16.3 squared meter and a range of 19.5 km. The nugget effect is variance at lag $h = 0$. Figure 3.6.3 shows the fitted model variogram model and its associated parameters. Using this fitted model, we may estimate the values of the variogram at any arbitrary spatial lag.

Variable	Year	n	Min	1st Qu	Median	Mean	3rd Qu	Max
GWL (m)	2009	526	0.3	3.5	6.7	8.0	11.3	46.7
RFL (m)	2009	526	0.4	0.5	0.7	0.7	0.8	1.1

Table 3.6.1: Summary Statistics, Groundwater Level (m) Post-Monsoon (Aug-Sep) and Total Monsoon (May, June, July, Aug) Rainfall (m) Western Uttar Pradesh, 2009

4. Step 5: Cross-Validation of the Fitted Model
   The fitted model was cross-validated by removing one observation at a time and using the 40 nearest values. Figure 3.6.4 shows spatial plots of the residuals between the observed and kriged values and the scatter plot and best fit line of the predicted versus observed values. 75% of the residuals lie between -1.7 and 1.7 meters. The slope of the regression line between predicted and observed values is 0.7 and adjusted R-squared is 0.7. The observed data is highly correlated with the predicted values indicating a good variogram fit.

5. Step 6: Kriging Interpolation
   Figure 3.6.5 shows the kriged groundwater level data and the prediction variance for 2009. White dots indicate the location of the wells. Areas of high depth to water table surround the urban centers of Meerut, Gautam Buddha Nagar, Baghpath, Bulandshahar and Agra whereas a groundwater rich region exists to the North-Eastern part of the study area in the Tarai belt. Prediction variance is high in the central parts where there are relatively fewer wells and consequently less available data.

3.6.3 The Impact of Spatial Auto-correlation on OLS Estimation

We use the simulation framework outlined in section 3.5 to assess the impact of spatial auto-correlation on OLS estimates in a simple regression of groundwater levels on rainfall. We use data for the year 2009 for a selection of districts within Western UP for this exercise. The data contains 564 well observations and has two variables, the Monsoon rainfall in 2009 within 5 km of well i denoted by X_i and the post-Monsoon groundwater level (measured at the end of August or so) denoted by G_i and the spatial coordinates of each well. Table 3.6.2 lists summary statistics and Figure 3.6.6 shows histograms of rainfall and GWLs data (top-left and top-right panels respectively), a scatter plot and regression line of GWL versus rainfall (bottom-right panel) and the locations of GWL observation wells (bottom-left panel).

We first obtain OLS estimates using real-world data, denoted by $\hat{\beta}^{rw}_{ols}$. Table 3.6.3

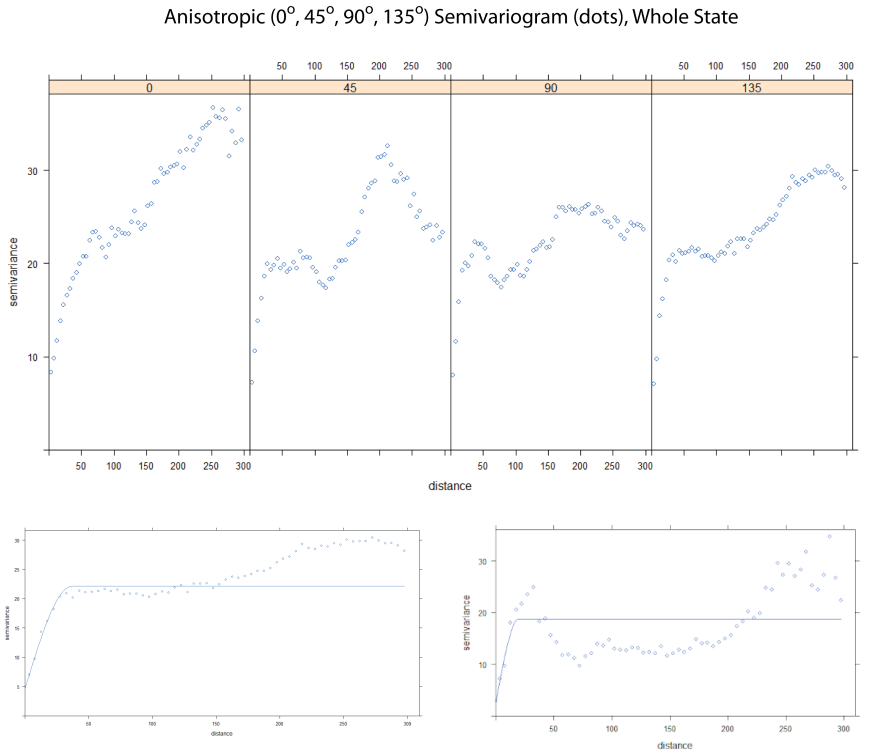
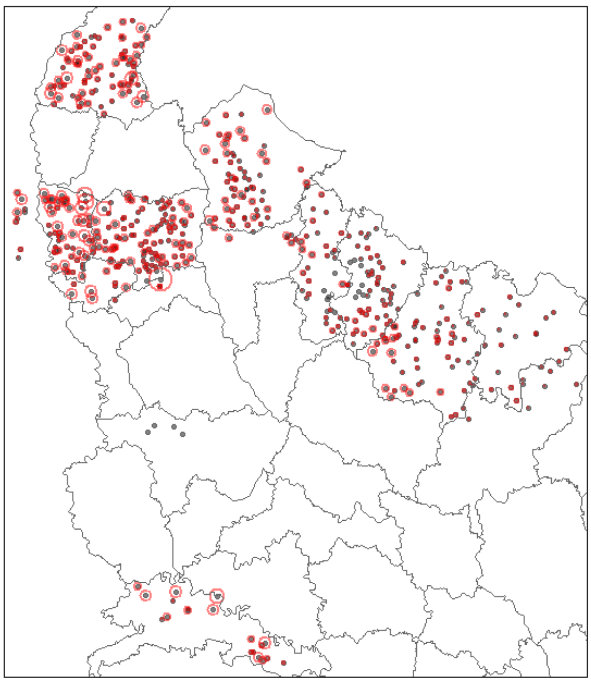
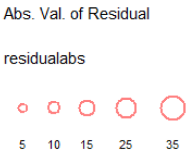


Figure 3.6.3: Top: Directional Semivariograms in the 0-135 degree directions for the entire state of UP. Bottom Left: 135 degree semivariogram and spherical theoretical variogram for the whole state of UP, width = 5 km, cutoff = 300 km. Range: 37 km, Partial Sill: 18 m^2, Nugget: 5 m^2. Bottom Right: 135 degree semivariogram and spherical theoretical variogram for four selected agro-climatic zones of UP, width = 5 km, cutoff = 300 km. Range: 20 km, Partial Sill: 16 m^2, Nugget: 2 m^2

Table 3.6.2: Summary Statistics

Variable	N	Mean	St. Dev.	Min	Max
Post-Monsoon GWL (m)	564	8.7	5.8	0.4	32.4
Rainfall (mm)	564	676.3	170.8	384.6	1,051.8
Latitude (dec deg)	564	29.0	0.5	28.0	30.3
Longitude (dec deg)	564	78.3	0.87	77.0	80.3



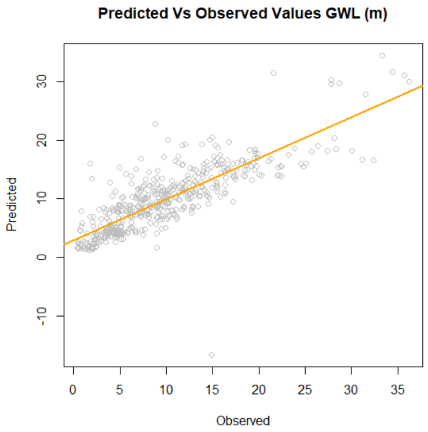


Variogram Model Cross-Validation

Left: Spatial Distribution of Residuals

Bottom Left: Predicted Vs Observed Values

Bottom Right: Regression Predicted Vs Observed Values



	Dependent variable:
	var1.pred
observed	0.701***
	(0.022)
Constant	2.913***
	(0.257)
Observations	506
R^2	0.670
Adjusted R^2	0.669
Residual Std. Error	3.224 (df = 504)
F Statistic	1,022.438*** (df = 1; 504)
Note:	*p<0.1; **p<0.05; ***p<0.01

Figure 3.6.4: Variogram Model Cross-Validation. Top: Spatial Distribution of Residuals. Bottom Left: Predicted Vs Observed Values. Bottom Right: Regression Predicted Vs Observed Values

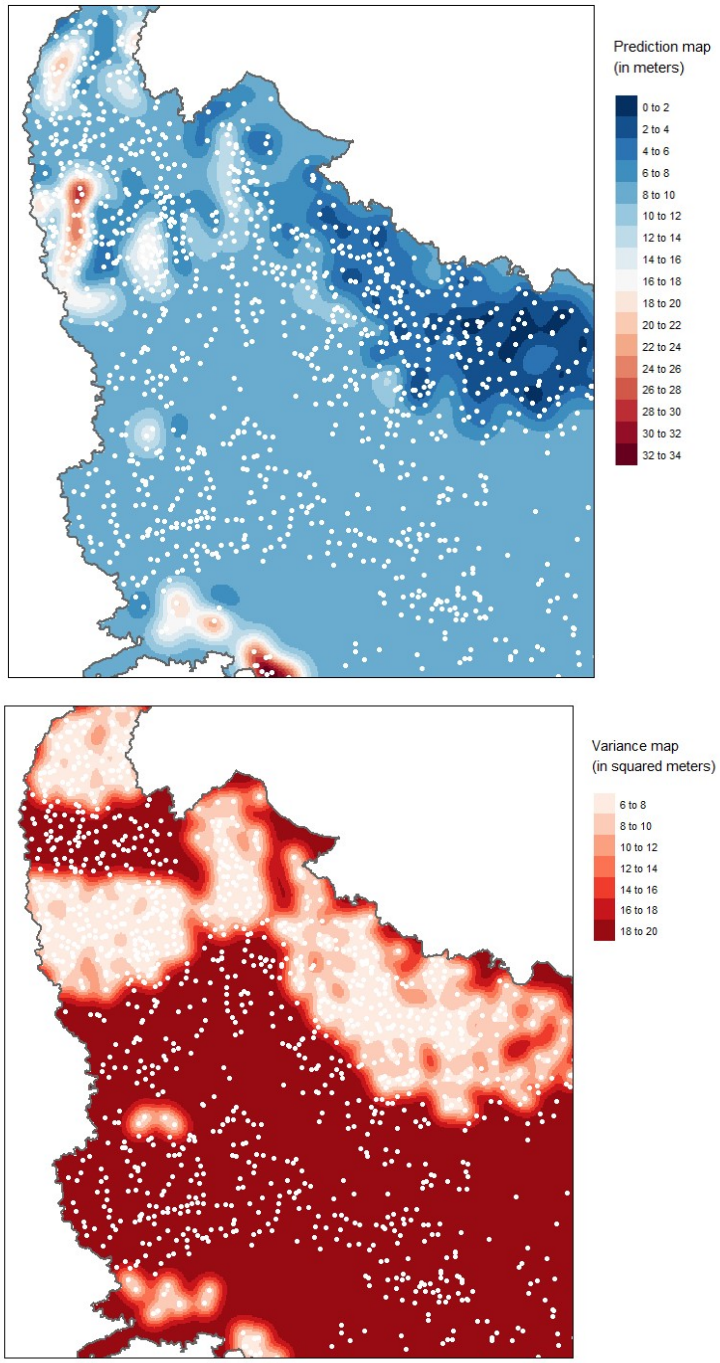


Figure 3.6.5: Kriged groundwater levels for 2009 (top) and the corresponding prediction variance map (bottom). White dots show observation well locations.

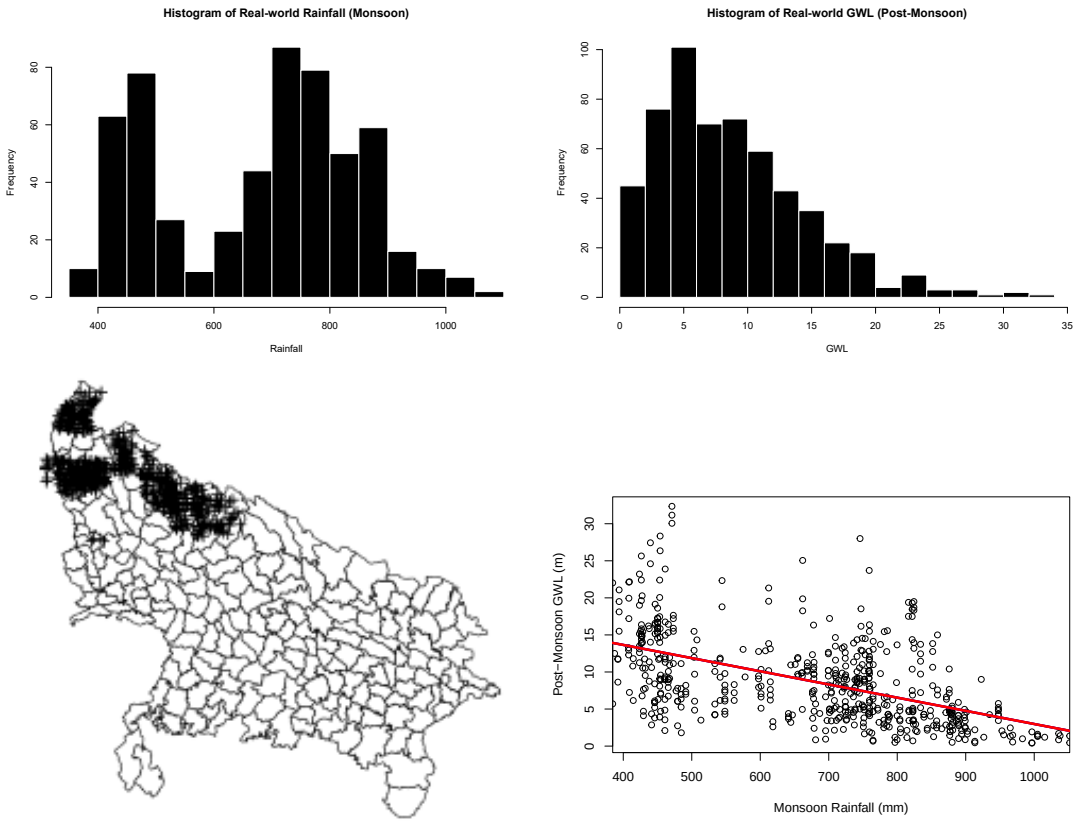


Figure 3.6.6: Monsoon Rainfall and Post-Monsoon GWL, Western UP, 2009. Real-world data. Top-left: Histogram of rainfall, top-right: histogram of GWL, bottom-left: locations of observation wells, bottom-right: scatter plot, GWL Vs Rainfall.

and Figure 3.6.6 (bottom-right panel) show the obtained results. We find that $\hat{\beta}_{ols}^{rw}$ = (21, -0.018). The estimate of the coefficient of rainfall is significant at the 99% confidence level. Monsoon rainfall explains close to 28% of the variation in GWL. An increase in rainfall by 1 mm is associated with a decrease (rise) of GWL by 18 mm. The estimated intercept is approximately 21 m. Using these exploratory regression results, we posit the true model as follows:

$$Y_i = 21 - 0.018X_i + \mu_i \tag{3.19}$$

where μ_i is a spatially autocorrelated error term with a spatial structure represented by a spherical variogram and sill (global variance) 33.

Table 3.6.3: Regression of real-world post-Monsoon GWL on Monsoon Rainfall.

	Dependent variable:
	Post-Monsoon GWL
Monsoon Rainfall	-0.018^{***}
	(0.001)
Constant	20.762^{***}
	(0.848)
Observations	564
R^2	0.276
Adjusted R^2	0.274
Residual Std. Error	4.928 (df = 562)
F Statistic	213.806^{***} (df = 1; 562)
Note:	*p<0.1; **p<0.05; ***p<0.01

We now simulate two sets of errors; first normally distributed errors that are distributed as ($u \sim N(0, 33I)$) where the number 33 is the variance of the real world GWLs. Second, spatially autocorrelated errors in which the true underlying spatial autocorrelation structure is given by a spherical variogram model with variance 33 m which is approximately the same as the sample variance of the GWLs data (SD = 5.8, see Table 3.6.2). Having simulated the errors, we compute synthetic data for GWL using the posited true model. Note that we do not use the real world GWL data directly because the purpose of this exercise is to assess the validity of variogram-based

GLS estimation and we would like to factor out variogram estimation error. We thus use a spatial autocorrelation structure of errors that is known apriori. An artifact of using synthetic data for GWLs instead of using the real-world data directly is that we can encounter occasional negative values. Negative values have no physical interpretation since GWL is measured from the ground surface which corresponds to a "depth" of 0 meters. GWLs are thus strictly positive. Other than the fact that they cannot be interpreted physically, negative values do not affect the statistical results or analytical premise of our simulation. We thus proceed with negative values rather than using truncation or other means to get rid of them.

Error Structure	$E[\hat{\beta}_{OLS}]$	$Var[\hat{\beta}_{OLS}]$	$E[\hat{\beta}_{GLS}]$	$Var[\hat{\beta}_{GLS}]$
Real World	-0.018	-	-	-
Simulated i.i.d	-0.018	0.0003	-	-
Simulated Sp. Auto-Correlated	-0.018	0.03	-0.018	0.023

SIMULATION PARAMETERS: MC iterations = 500, Samples = 564, True β = (21, -0.018), True Variogram Model = Spherical, True Sill = 33, True σ^2 = 33

Table 3.6.4: Monte Carlo Simulation Results

The results from 500 Monte Carlo iterations are summarized in Table 3.6.4. The real-world estimate was -0.018 which was also set as the true model parameter. Both OLS and GLS produced unbiased estimates of -0.018 that equal to the true parameter value irrespective of the error structure. The introduction of spatially auto-correlated error structure caused the variance of OLS (column 2) to increase 10 times from 0.0003 in the case of i.i.d errors (row 2) to 0.03 (row 3). The increase in variance is consistent with theory in that spatial auto-correlation causes the OLS estimates to remain unbiased but renders them inefficient (N. A. Cressie, 1993; Beale et al., 2010).

We estimated the regression using GLS with a covariance matrix obtained from the fitted variogram. This also led to results that accord with theory. The point estimate (-0.018) remained unbiased and the variance decreased from 0.03 to 0.023 thus providing a tighter estimate.

Figure 3.6.7 visualizes the results from the table. The right-panel shows the variance of the estimators and we see that all variances are significantly different from each other with non-overlapping 95% confidence intervals. The variance of the GLS estimator is

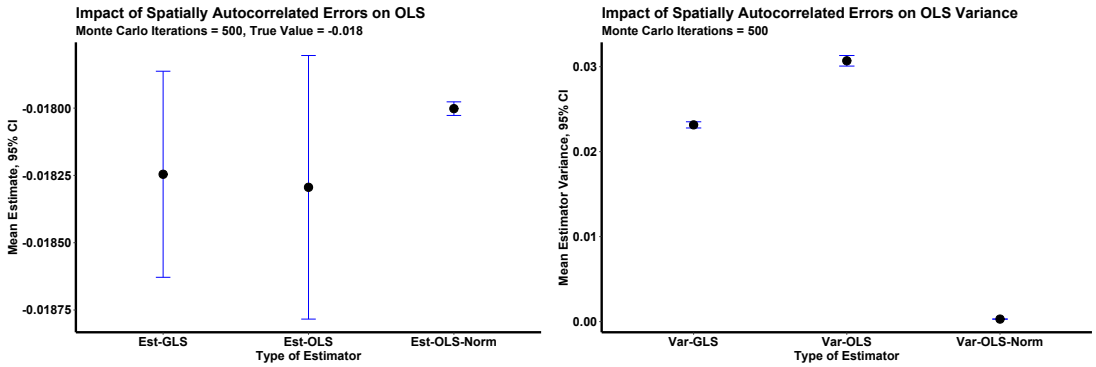


Figure 3.6.7: Estimates (left panel) and Variances of Estimates (right panel) for OLS with normally distributed i.i.d errors (OLS-Norm), OLS with spatially autocorrelated errors (OLS) and GLS with spatially autocorrelated errors (GLS).

significantly lower than the OLS estimator. The inde of "loss of efficiency" when the OLS estimator is used instead of the GLS estimator is defined as follows by Doran and Griffiths (Doran & Griffiths, 1983):

$$\alpha_{re} = \frac{|Var[\hat{\beta}_{OLS}]|}{|Var[\hat{\beta}_{GLS}]|} - 1 \tag{3.20}$$

In our case, the index of loss of efficiency is $\alpha_{re} = \frac{0.03}{0.023} - 1 = 0.3$ equivalent to a 30% loss. Figure 3.6.8 depicts this loss visually. It shows the densities of the OLS and GLS estimators based on 500 estimates from our simulation. It can be seen that while the mean value of both estimators is the same (dashed lines) GLS provides a relatively tighter estimate.

3.7 Conclusion

We have discussed the concept of spatial auto-correlation and provided a theoretical background on the geo-statistical techniques applied in capturing the spatial structure in groundwater level data. Further, we apply these techniques to extract an experimental variogram using governmental groundwater levels data from Western Uttar Pradesh in Northern India. We find suggestive evidence that groundwater levels in the region exhibit positive spatial auto-correlation. This may reduce the precision of OLS estimates on account of violation of the classical assumption of independent errors needed

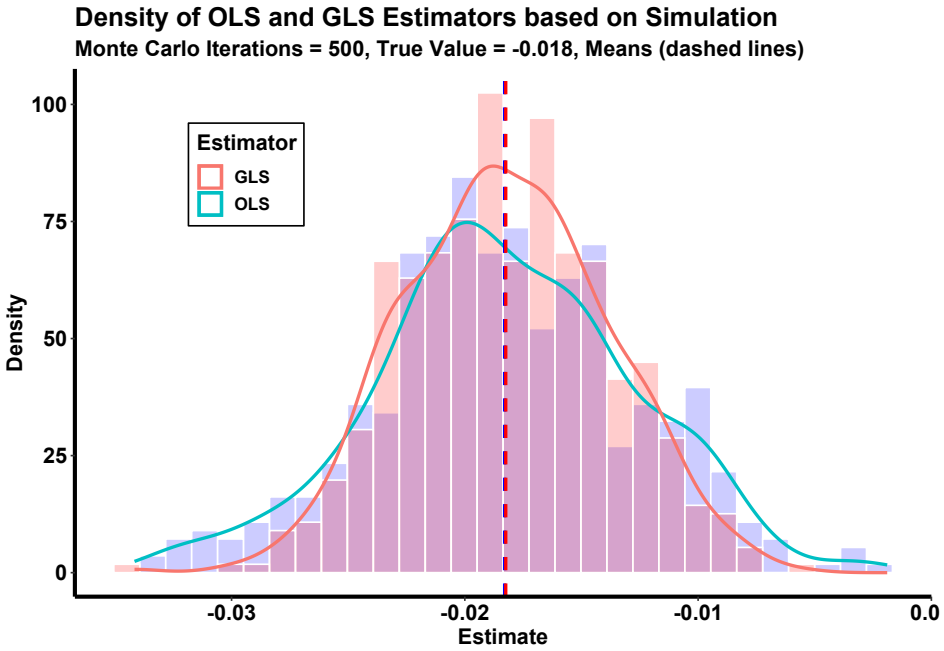


Figure 3.6.8: Density of the OLS (blue) and GLS (red) estimators.

for minimum variance estimates according to the Gauss Markov theorem. We present a method to overcome this limitation by the use of feasible generalized least squares and demonstrate how the variance-covariance matrix of the error terms can be recovered from a fitted variogram model. We then present a simulation study that assesses the impact of spatial auto-correlation on OLS estimates and the efficacy of our variogram-based fGLS method in providing relatively more efficient estimates. Our results corroborate the loss of efficiency of OLS estimates when spatial auto-correlation is present. They also theoretically show that if variograms are estimated correctly, fGLS can be used to regain precision of least squares estimates of the conditional mean of GWL.

While these results suggest that variogram-based GLS estimation offers a potential way out of the statistical uncertainty induced by spatial autocorrelation in GWLs, we have not addressed the variogram estimation error. Our results were obtained with a simulated and known spatial structure. In future work, we plan to explore rigorous hypothesis testing and goodness of fit measures for variogram estimation which will enable us to estimate the underlying spatial structure of GWLs in various aquifers and geographical settings and also provide confidence intervals for such estimation.

CHAPTER 4

GROUNDWATER LEVELS DYNAMICS AND AGRICULTURAL LAND-USE CHANGE IN THE GANGETIC BASIN

4.1 Introduction

The goal of this study is the econometric estimation of spatially-delineated groundwater level (GWLs) dynamics in relation to agricultural land use intensification. The study area is the Northern Indian state of Uttar Pradesh (UP) and the study period is 2009-10 to 2018-2019. Identification of land use impacts on groundwater will rely on a stark shift in agricultural land use changes in the part of the Gangetic Basin (GB) covering UP due to an increase in multiple cropping acreage during the study period (Arora et al., 2019). Such land use changes are driven, at least partially by, access to groundwater irrigation supported by conducive policy interventions like highly subsidized electricity and distribution of public tube wells (Pant, 1994). Significant increase in minimum support price (i.e., floor price) for crops procured by the government and large-scale agricultural loan waivers circa 2009 are also potential drivers of land use intensification in the GB. We will exploit exogenous variation in land use changes through an instrumental variables regression framework to ultimately establish the causal impact of land use changes on GWLs.

The GB is a major source of national agricultural production including for staples like wheat, rice and pulses, and also cash crops like sugarcane and maize. The GB is endowed with a fertile soil cover and thick alluvial aquifers with high water bearing capacity and abundant Monsoon rainfall that increases from west to east (Muñoz Sabater, 2019) (see Figure 4.A.2). A multi-layered aquifer system underlies UP forming one of the richest groundwater repositories of the world where groundwater contributes 71% of the total irrigation water needs of the state (CGWB, 2021). The state of UP is an

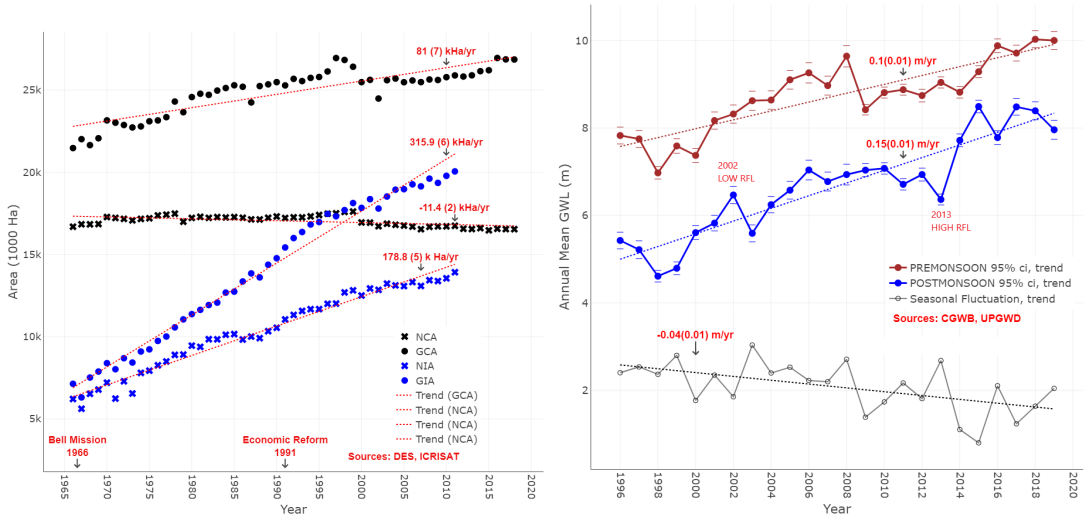


Figure 4.1.1: Left panel: Gross and Net Cropped and Irrigated Area, UP, 1966-2018, Right panel: Annual Mean GWLs, UP, 1996-2019

agrarian economy that was home to 16.5% of the total population of India in 2011 (Census of India, 2011) of which 47% directly depended on agriculture to make their living (Gulati et al., 2021).

Economists have highlighted that at the current (2018-2021) rate of withdrawals, groundwater resources may be limited in their capacity to sustain further agricultural expansion (Jain et al., 2021) or serve as a means of adaptation to rainfall variability in India (Fishman, 2018). A study reported that in the dry (winter) season, there was a 3% decline in mean rice yields at the national scale for every one meter depletion in local groundwater table during 2004–2013 (Bhattarai et al., 2021a). GWL in UP increased (deepened) at an average rate of 0.1 (PREMONSOON) to 0.15 (POSTMONSOON) meters per year between 1996 and 2019 (see Figure 4.1.1, right panel) with relatively greater increases in drought years (2002-03, 2015-16) and recoveries during good rainfall years (2003-04, 2013-14). The difference between the pre-monsoon and post-monsoon levels, called the "seasonal fluctuation", has decreased by 0.04 m every year on average since 1996 indicating a gradually reducing influence of monsoon rainfall in aquifer recharge. Closely spaced drought years 2012 and 2015 preceded the minimum seasonal fluctuation recorded in the period (0.8 m) in the year 2015 (Guhathakurta et al., 2020).

Agricultural land-use change in India in the 20th century was characterized by the expansion of net cultivated area in the first half of the century and the expansion of

multiple cropping in the second half of the century. The major expansion in net cropped area in the post-Independence years occurred in the 1950s, the fastest growth occurring between 1941 and 1961 (Kurosaki, 2007). In the 1960s and later years, very little new land was brought into cultivation and expansion of cultivated area was achieved via multiple cropping of existing agricultural lands (Abel, 1970) (see Figure 4.1.1, left panel). According to official land-use statistics, in the part of the GB covered by UP, gross cropped area (GCA) increased by approximately 20% between 1966 and 2000. After the year 2000, official statistics from UP report a relatively small (6%) increase in GCA whereas our estimates from high resolution remotely sensed data suggest an increase of the order of 50%; a huge deviation from official figures. For further detail, we present a brief historical account of land-use intensification in UP since the Mughal era to the post-independence era in Appendix 4.A.

There is evidence that groundwater irrigation was a strong driver of agricultural expansion after independence (1950s) up until the 1990s (Dayal, 1977; Narain & Roy, 1980; Dhawan & Datta, 1992) although some economists questioned such findings (Sawant, 1975). Between 1966 and 2012, gross irrigated area (GIA) increased by 150% with growth being particularly fast after economic reforms in the 1990s (DES, 2008; ICRISAT, 2015). The concerns around of groundwater depletion first emerged in the 1990s on account of declining GWLs (Chandrakanth & Romm, 1990). In the Northern Plains of the Indian subcontinent, the euphoria around groundwater development in the years after independence gave way to an urgent need for groundwater resource management in the years leading up to millennium (Shah, 2010; Subramanian, 2015; Dantwala, 1976).

In the 2000s, a large and influential research literature has highlighted that the agricultural sector is the largest user of groundwater in India (Anuraga et al., 2006; Rodell et al., 2009; Siebert et al., 2010; Scanlon et al., 2012; Wada et al., 2014; Ahmed & Umar, 2009; B.M.Jha, n.d.). An estimated 89% (217.61 billion cubic meter) of the total withdrawal in 2020 was for irrigation (CGWB, 2021). The paradigm shift from development to management has occurred on account of severe yet spatially non-uniform groundwater depletion across the sub-continent (CGWB, 2019; Rodell et al., 2009).

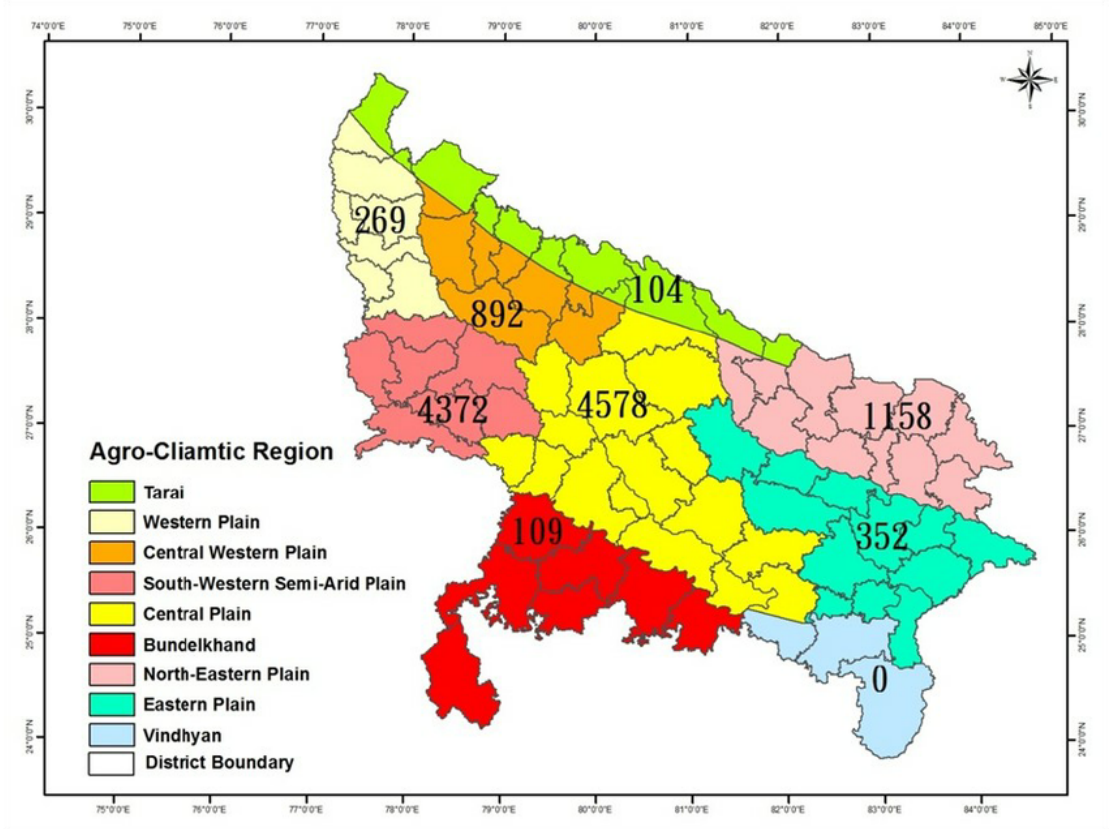


Figure 4.2.1: Study Area: Uttar Pradesh (UP), North India, Agro-climatic zones.

4.2 Background

We study agricultural land-use intensification in UP with the aim to estimate its causal impact on regional GWL dynamics. UP is located within the GB of India, spreading between 23° 50'–30° 45' N latitude and 77° 04'–84° 38' E longitude, as shown in Figure 4.2.1. It is divided into nine "agro-climatic zones". An agro-climatic zone is a spatially delineated region which is considered a single entity due to the similarity of geographical, climatic and agrarian conditions within its boundaries. Agro-climatic zonal planning was part of the National Agricultural Research Project (NARP) launched by the Indian Council of Agricultural Research (ICAR) in 1979, with World Bank assistance. Each state was divided into several agro-climatic zones to based on climate, soils, crops and ecology to strengthen *region-specific* planning and research (Kareemulla, 2022).

The boundaries of agro-climatic zones accord with cropping patterns in the state of UP (Figure 4.2.2, top-left panel). Sugarcane systems prevail in the Western Plain, maize and *jowar* (a type of coarse cereal) in the Southwestern semi-arid Plain, pulse systems

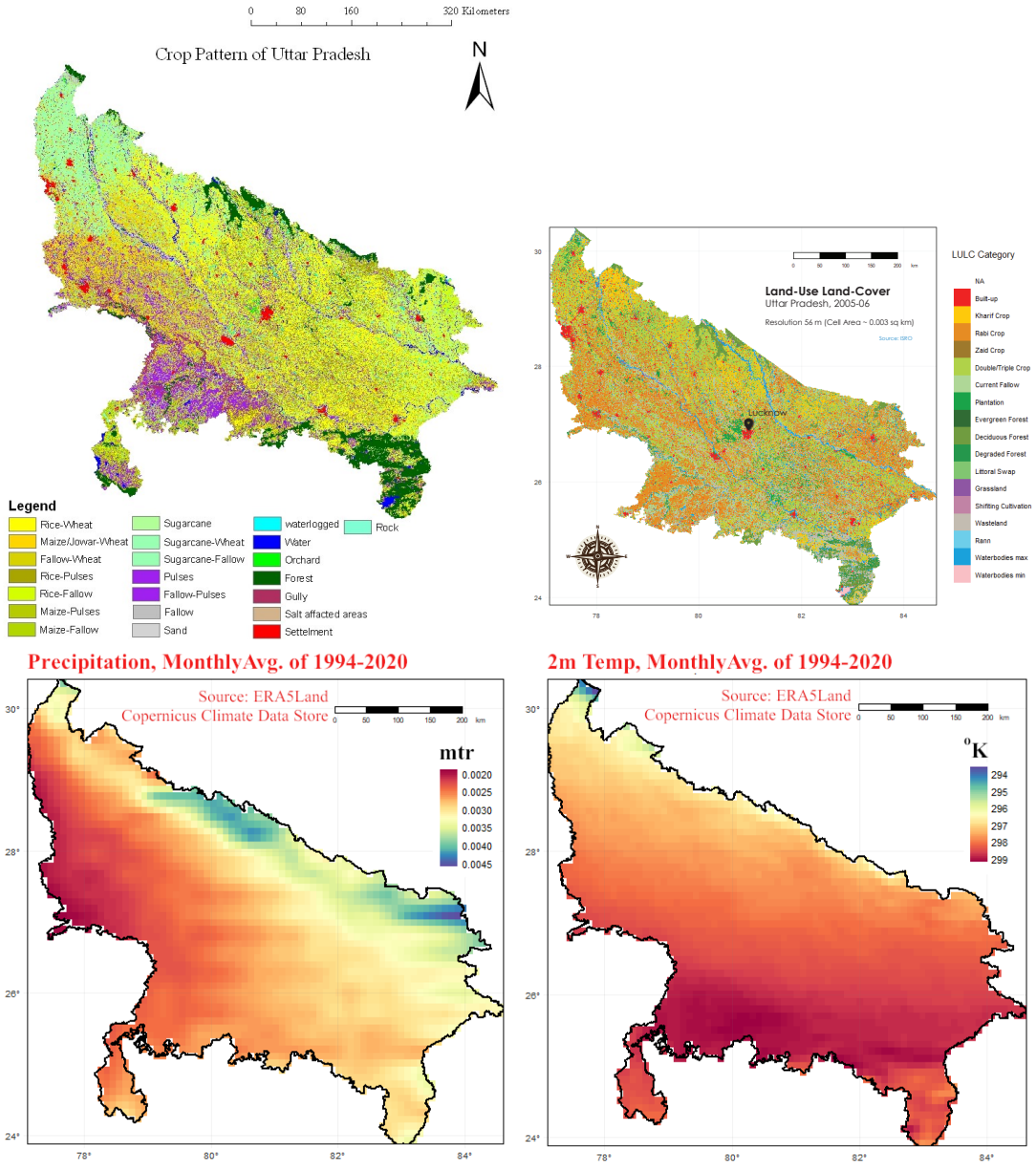


Figure 4.2.2: *Top-left*: Cropping pattern of Uttar Pradesh, 2009. Data synthesized from from remote-sensing data by Singh (Singh et al., 2011). *Top-right*: LULC categories map synthesized from remote-sensing data by the National Remote Sensing Center (NRSC, 2007). *Bottom-*: Mean monthly rainfall (*left*) and 2m temperature (*right*), 1994-2020, source: ERA5 Land dataset from the Copernicus Climate Data Store (Muñoz Sabater, 2019).

in Bundelkhand while rice and wheat systems dominate the Central Western, Central, Eastern and North-Eastern Plains. The Vindhyan zone is mostly forested area. Factors affecting GWL dynamics, i.e., hydrological attributes (aquifer structure and surface water sources), climatic factors (mild and dry winters and hot summers) and groundwater demand (irrigation and urbanization), are found to vary substantially across UP. Weather patterns are rerpesented by cool and dry weather (464 mm average annual rainfall) in the northern and western parts of the state while hot and wet weather (1,245 mm average annual rainfall) in the southern and eastern parts of the state (see Figure 4.2.2, bottom left and bottom right panels).

We measure land-use transitions from 2009-10 to 2018-19 using high resolution images from remote-sensing satellites that indicate land-use land-cover (LULC) categories for each year and the entire region at a spatial resolution of 56 meters. Each "cell" or pixel in a raster image is classified into one of 18 categories (see Figure 4.2.2, top-right panel for an example of the LULC raster for 2005-06). Each cell occupies an area of 56m $\times$ 56m (approximately 0.003 sq km.) on the ground (NRSC, 2007).

A multi-layered alluvial aquifer system and dense network of rivers and canals along with abundant annual rainfall of 784 milimeters (mm) together form one of the richest groundwater repositories in the world. However, there exists a west to east trend in GWLs - up to 30 meters (m) deep in the western districts while less than 3 m in the eastern districts (Figure 4.2.2, bottom-left panel). Each cell in the raster contains a value between 0 and 18 indicating a

The state of UP has three major cropping seasons, *kharif* (rainy season, June to October), *rabi* (dry winter season, October to April) and *zaid* (dry summer season, April to June) (see Figure 4.3.1. While irrigation requirements for in *rabi* and *zaid* crops are higher, they are not negligible for *kharif* crops. A study has reported that the percentage of groundwater irrigation (of the total amount of irrigation water) in India as a whole is 79% in Rabi, 63% in Zaid, 27% in Kharif (Biemans et al., 2016). It is thus very likely that land-use transitions from single cropping in any one season in a year to multiple cropping in two or more seasons in the next year will entail a rise in irrigation water use. However, it is expected that transitions from *kharif* to multiple cropping signal a greater increase in irrigation water use as compared to transitions from *rabi* or

Type	From	To
HI	*kharif* (2)	double-triple crop (5)
MI	*rabi* (3)	double-triple crop (5)
MI	*zaid* (4)	double-triple crop (5)

Table 4.2.1: Land-use transitions and irrigation requirements

zaid to multiple cropping.

We thus consider two different types of land-use transitions. We term these "moderate-irrigation transitions" (MI) and "high-irrigation transitinos" (HI). Transitions from *kharif* (LULC category 2) to "double-triple cropped" (category 5) are HI transitions and transitions from *rabi* (LULC category3) or *zaid* (LULC category 4) to "double-triple cropped" (category 5) are MI transitions. Table 4.2.1 lists each type of transition and specifies the initial and final land-use categories associated with it.

4.3 Materials and Methods

The groundwater balance or groundwater budget is a measurement of continuity of the flow of water in a given time interval through a groundwater system. The systems usually considered by hydrologists are open hydro-geological units like aquifers, watersheds, and drainage basins. Groundwater flow and dynamics are studied within spatially delineated volumteric regions called *control volumes*. A control volume is a specific region of space within a groundwater system over which groundwater balance equations are defined. The quantification of the hydrologic cycle is a mass balance (equation 4.1) in which the change in water storage with respect to time $\frac{dS}{dt}$ within the considered system is equal to the inputs I to the system minus the outputs O from the system (Todd & Mays, 2005).

$$\frac{dS}{dt} = I - O \tag{4.1}$$

The water balance equation is useful for studying the response of hydrologic systems to climatic processes and human stressors. The components included in the equation must be tailored to each specific region and application. Wieskel has adapted the

basic water balance to the form given in 4.2 to study the direct human interaction with aquifers (Weiskel et al., 2007).

$$\frac{dS}{dt} = (R_p - D_{et}) + (R_{gw} + R_{sw}) + H_{in} - (D_{gw} + D_{sw}) + H_{out} \tag{4.2}$$

R_p is aquifer recharge from precipitation; R_{gw} and R_{sw} are aquifer recharge from adjacent groundwater and surface water systems, respectively; D_{et} is groundwater evapotranspiration; D_{gw} and D_{sw} are aquifer discharge to adjacent groundwater and surface water systems; H_{in} is total return flow to the aquifer; H_{out} is aquifer withdrawals; and $\frac{dS}{dt}$ is the rate of change in aquifer storage. All units are length^3/time (L^3/T) averaged over the period of interest. All flow terms are positive,except $\frac{dS}{dt}$, which can be positive, negative or zero during the period of interest (Weiskel et al., 2007). The change in storage over a time interval is related to the water table fluctuation through 4.3 where ΔS is the total change in the volume of water stored; Δg is the change in water level over the given time interval, A is the cross-section of the control volume and Φ is the *specific yield* of the aquifer control volume (Maréchal et al., 2003). Specific yield is defined as the ratio of the volume of water that a saturated rock or soil will yield by gravity to the total volume of the rock or soil. Specific yield is usually expressed as a percentage (Johnson, 1963).

$$\Delta S = \Delta g.\Phi.A \tag{4.3}$$

The area A corresponds to the cross-sectional area of the control volume. The term Δg is the fluctuation of GWLs over a period of time (annual, decadal etc.) aggregated over the area A[1]. We can rearrange equation 4.3 to bring the change (fluctuation) in GWL to the left hand side and denote the change over a time interval by the subscript t over a control volume i as follows:

$$\Delta g_{i,t} = \frac{\Delta S_{i,t}}{\Phi_i A_i} \tag{4.4}$$

[1]The equation 4.3 is the equation used in computing seasonal changes in storage by the CGWB by the "water table fluctuation" method. GWL fluctuations are measured at monitoring wells and spatially aggregated to compute the fluctuation across a control volume.

Note that specific yield and cross-sectional area are constants over time for a given control volume. We can expand the RHS of equation 4.4 using equation 4.1.

$$\Delta g_{i,t} = \frac{(R_{p,i,t} + R_{gw,i,t} + R_{sw,i,t} - D_{et,i,t} - D_{gw,i,t} - D_{sw,i,t} + H_{in,i,t} - H_{out,i,t})}{\Phi_i \times A_i} \tag{4.5}$$

Equation 4.5 expresses the structural relationship between change in GWL and the change in water storage within a control volume i over a time period t. Each term in 4.5 is a deterministic variable that represents a volume of water either entering (recharge) or leaving (discharge or extraction) the aquifer control volume. For some terms like recharge due to rainfall $R_{p,i,t}$, we can obtain data that provides a direct measure but for others like recharge (or discharge) due to surface water bodies $R_{sw,i,t}(D_{sw,i,t})$, data is not available and we must use proxy variables. The use of proxies introduces uncertainty since they provide correlations with but not exact measures of each variable. Further, one proxy variable may be correlated with multiple terms in the equation. For example, the proxy "distance to nearest surface water body" is correlated with both the recharge from surface water bodies $R_{sw,i,t}$ and the discharge from surface water bodies $D_{sw,i,t}$.

Due to the uncertainty introduced by inexact correspondence between available data and the deterministic terms of the groundwater balance equation 4.5, we replace the deterministic terms by random variables and use a linear regression model to perform estimation. Our goal is to estimate the causal impact of agricultural land-use intensification on GWLs.

Unit of Analysis and Dependent Variable: The unit of analysis i is a monitoring well. The dependent variable $\Delta g_{i,t}$ is the annual change in GWL inside the monitoring well i between May of year $t-1$ and year t as given by equation 4.6.

$$\Delta g_{i,t} = g_{prem,i,t} - g_{prem,i,t-1} \tag{4.6}$$

where $g_{prem,i,t}$ is the pre-monsoon (May) GWL at well i in May of year t. Figure 4.3.1 shows a schematic timeline that we term a "hydrologic schedule." It show the de-

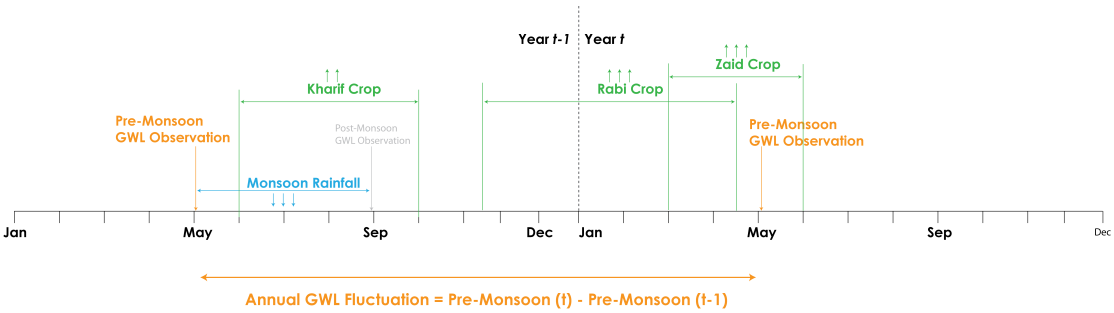


Figure 4.3.1: Annual change in GWL, Cropping Seasons and Rainfall

pendent variable, i.e the difference between pre-Monsoon GWL observations are made in May of calendar year $t-1$ and calendar year t, as well as the occurrence of the Monsoon rainfall and the major cropping seasons, *kharif* (rainy season, June to October), *rabi* (dry winter season, October to April) and *zaid* (dry summer season, April to June), for the UP region.

Covariates: We construct a set of spatially delineated covariates withing "spatio-temporal neighborhoods" for each well. If a GWL observation g_i was made in year t at well location (x, y) then the spatio-temporal neighborhood of this observation is a set of locations and time periods that satisfy the following condition.

$$N_{s,i} \; = \; \{ \, (x', y', t') : \; s.t \, |(x', y') - (x, y)| \leq r \; and \; t' = t - l \, \} \tag{4.7}$$

where r is called the "radius" and l the "lag" of the spatio-temporal neighborhood and $|.|$ is the Euclidean distance operator. Note that the lagged time period t' can only be in the past thus we must have $l > 0$ and $t' < t$. For the "current" monitoring year t, the lag is zero.

Land-use transitions: The variables of interest are percentage area that experienced HI transitions ($HI_{i,t}$) and MI transitions ($MI_{i,t}$) in the spatio-temporal neighborhood of well i from year $t-1$ to t.

Climate: Covariates that capture climatic factors include Monsoon precipitation in year t - $p_{i,t}$ and average 2 meter temperature in the summer (Apr, May) in year t - $t2m_{i,t}$.

Geography/Hydro-geology: We control for soil quality with a measure of the soil organic carbon density around well i - sc_i. Sub-surface recharges and discharges between rivers and aquifers influence GWL dynamics. We control for this process using the dis-

tance of well i to the nearest water river dr_i[2]. We also control for the aquifer type in which well i is located - aq_i. We use a dummy variable with values corresponding to three different aquifer types according to a map of aquifer systems in UP prepared by Bhanja (2019). The rate of sub-surface flow and storage capacities vary across aquifer types and this influences GWL dynamics.

Infrastructure: Land-use transitions toward multiple cropping are likely to be greater in areas where transport routes for agricultural commodities are nearby. We quantify the availability of transport in the proximity of a well by the distance to nearest national highway $dnh_{i,t}$.

Urban demand: Urban demand for agricultural commodities can drive land-use transition toward multiple cropping in the neighborhood. We selected 9 largest urban centers in Uttar Pradesh by population. We use the distance of a monitoring well to the nearest urban center $dub_{i,t}$ to measure the influence of urban demand. We also control for groundwater extraction for domestic use with the total population of habitations within the spatial neighborhood of each well - $pop_{i,t}$. **Policy**: The CGWB classifies each block into one of four categories according to the stage of local groundwater development in that block (CGWB, 2019). We include a dummy variable which indicates the CGWB block classification of the block in which well i is located - $cl_{i,t}$. The classes are *safe* (0), *semi-critical* (1), *critical* (2) and *over-exploited* (4). The CGWB assigned new classes to blocks in 2011 and 2013.

The model is specified by the following equation:

$$\Delta g_{i,t} = \beta_0 + \eta_1 HI_{i,t} + \eta_2 MI_{i,t} + \boldsymbol{X}\boldsymbol{\beta} + \epsilon_{i,t} \tag{4.8}$$

where $\boldsymbol{X}$ contains the covariates and $\epsilon_{i,t}$ accounts for the unobserved sources of GWL fluctuation. We are interested in the estimates $\hat{\eta_1}$ and $\hat{\eta_2}$ that measures the impact of irrigation intensive agricultural land-use transitions on the change in GWLs.

Instruments: Estimation of $\hat{\eta_1}$ and $\hat{\eta_2}$ (from equation 4.8) using ordinary least squares (OLS) is likely to be biased due to the endogeneity of land-use transition to GWL. The Figure 4.3.2 shows a directed acyclic graph (DAG) of the dependencies be-

[2]We restricted the set of rivers to those that have a Strahler number greater than or equal to 3. This excludes minor streams.

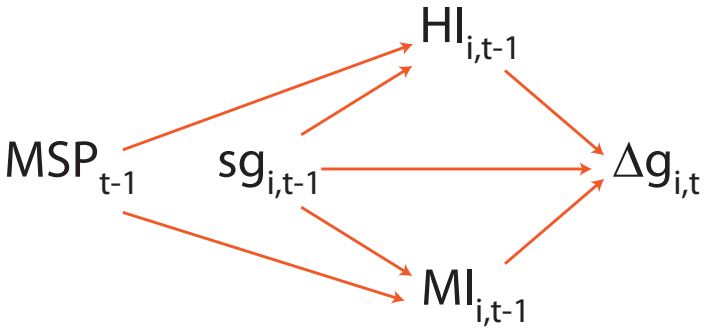


Figure 4.3.2: Directed acyclic graph showing dependencies between variables. An arrow indicates that the variable on the tail of it has direct influence on the variable at the head.

tween the different variables in our specification. An arrow indicates that the variable at the tail has a direct influence on the variable at the head. The dependent variable is $\Delta g_{i,t}$; the change in GWL at a well location i between years $t-1$ and t. This is directly influenced by the irrigation decisions during the year $t-1$ in the spatial neighborhood of well location i. We thus have arrows from the independent variables $HI_{i,t-1}$ and $MI_{i,t-1}$ to $\Delta g_{i,t}$.

Endogeneity arises from the fact that both the irrigation decisions and the change in GWL in a particular year is influenced by the local stage of groundwater development which we denote as $sg_{i,t-1}$. In India, the SGE is the indicator of the local stage of groundwater development and it depends on groundwater recharge and utilization in the past year(s) and long-term GWL level trends in the region (CGWB, 2021). This directly influences irrigation decisions since the cost of groundwater extraction in a region of high SGE is greater than one of relatively lower SGE. It also directly influences the change in GWL at well location i in the next year t thus it is a confounding variable. Areas of higher groundwater depletion tend to occur in places where SGE is also high and thus it is not generally possible to disentangle the impact of irrigation.

We employ a 2-stage least squares (2SLS) approach using the minimum support price (MSP) as an instrument to correct for endogeneity. The MSP is a nationwide floor price exogenously assigned for 28 agricultural commodities by the Commission for Agricultural Costs and Prices (CACP).

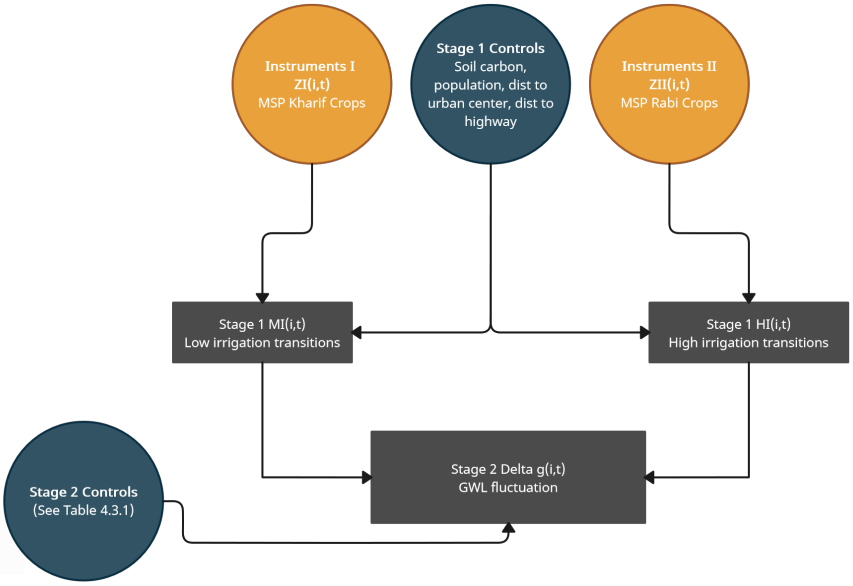


Figure 4.3.3: Dependent variables, controls and instruments in the 2SLS setup.

In a given cropping year, the government guarantees procurement of each agricultural commodity from farmers at the pre-specified MSP level. Government procurement of crops at the MSP price level is shown to have been a driver of agricultural land-use intensification in Punjab (Pandey, Khan, & Kumari, 2019). Farmers, assured by a price floor, are driven to invest in irrigation infrastructure to transition landholdings from low or no-irrigation to high and moderate irrigation cropping patterns. We thus expect the instrument to be relevant implying that $Cov[MSP_{t-1}, HI_{i,t-1}] \neq 0$ and $Cov[MSP_{t-1}, MI_{i,t-1}] \neq 0$.

The key to the exogeneity of the MSP instrument is the fact that it only varies by year and not spatial location. The CACP considers a number of factors including the demand and supply of each commodity and the cost of production which are influenced by the cost of groundwater irrigation but sets a single price for each commodity across the country (CACP, 2022). The MSP for each commodity is thus exogenous to the regional groundwater scenario (ISEC, 2003). The price, for a given commodity, will be the same in two regions even if they differ in the extent of local groundwater development. We thus expect that $Cov[MSP_{t-1}, sg_{i,t-1}] = 0$. Naturally, the MSP does not directly influence the GWL but only through land-use transitions to intensive irrigation regimes thus we expect that $Cov[MSP_{i,t-1}, \Delta g_{i,t}] = 0$.

We construct 6 instrumental variables $msp^{l}_{i,t}$ where $l = 1, ..., 6$ using public data

published by the CACP (CACP, 2022). Farmers observe the MSP for each crop usually before the sowing time in each season and decide the crop mix for that season. We use as our instruments the MSP per unit of the top three crops by area allocation in each of the two major cropping seasons (*kharif* and *rabi*) that were grown in the district where the well i is located. Even though MSPs are declared for each year, in our model they vary at district level since districts have different cropping systems. We do not have well-level variation in the instrument. This spatial resolution constraint is imposed by district-level data availability for crop-wise area. Appendix 2.D describes each covariate and the data construction process in detail. Table 4.3.1 provides summary statistics.

In the first stage of our 2SLS setup, we have two regressions. We regress $MI_{i,t}$ on a set of instruments and controls $\boldsymbol{Z}^{I}_{i,t}$ and separately regress $HI_{i,t}$ a another set of instruments and controls $\boldsymbol{Z}^{II}_{i,t}$.

$$MI_{i,t} = \alpha_0 + \alpha_1 Z^{I}_{i,t} + u_{i,t} \tag{4.9}$$

$$HI_{i,t} = \gamma_0 + \gamma_1 Z^{II}_{i,t} + v_{i,t} \tag{4.10}$$

In the above specifications, $Z^{I}_{i,t}$ contains the three instrumental variables and four control variables. The instruments are the MSPs of the top three *kharif* crops grown in year t within the district in which the well i is located. The control variables are $\{dnh_{i,t}, dub_{i,t}, sc_{i,t}, pop_{i,t}, cl_{i,t}\}$ (see Table 4.3.1 for details) and $u_{i,t}$ is the unobserved error term.

$Z^{II}_{i,t}$ contains the three instrumental variables and four control variables. The instruments are the MSPs of the top three *rabi* crops grown in year t within the district in which the well i is located. The control variables are $\{dnh_{i,t}, dub_{i,t}, sc_{i,t}, pop_{i,t}, cl_{i,t}\}$ (see Table 4.3.1 for details) and $v_{i,t}$ is the unobserved error term.

In the second stage, we regress $\Delta g_{i,t}$ on estimates obtained in the first stage along with a set of controls $\boldsymbol{X}$. The setup is specified below:

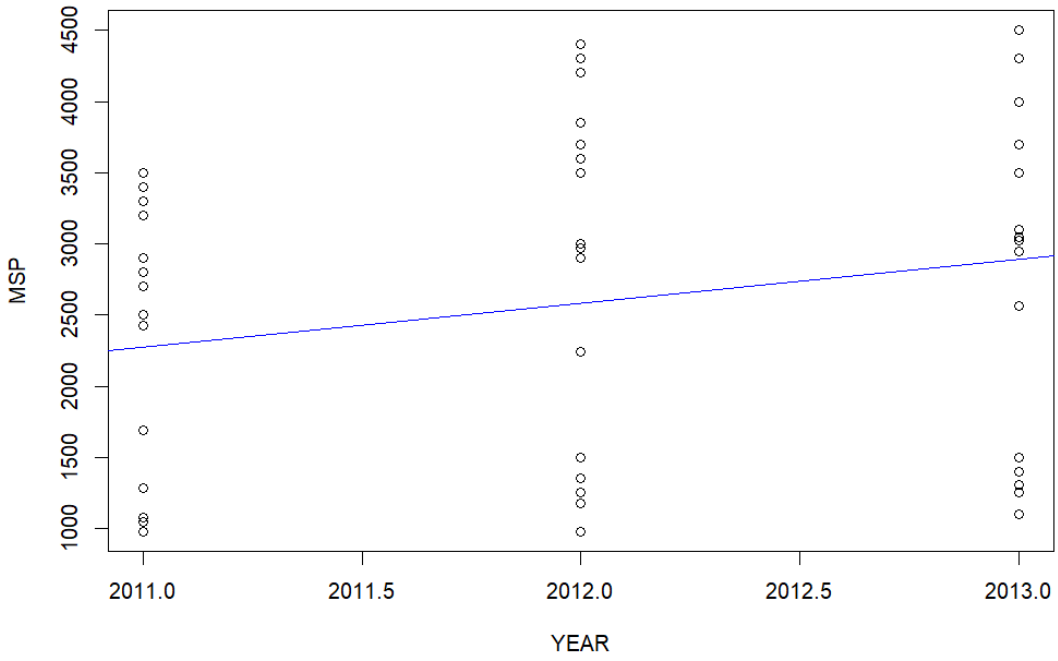


Figure 4.3.4: MSPs for crops in UP, 2011-12 to 2013-14.

$$\Delta g_{i,t} = \beta_0 + \eta_1 \hat{LI}_{i,t} + \eta_2 \hat{HI}_{i,t} + \beta_1 X_{i,t} + \epsilon_{i,t} \tag{4.11}$$

The figure 4.3.3 illustrates the 2SLS model specification.

4.4 Preliminary Results

We estimated $\hat{\eta}_1, \hat{\eta}_2$ with a two-stage least squares model with data pooled over years. Table 4.3.1 provides summary statistics for the dependent, independent and instrumental variables as well as the controls. Figure 4.4.1 shows the spatial pattern of HI and MI transitions and annual change in GWLs between 2011-12 and 2013-14. It is clear that HI and MI land-use transitions represent two different regimes of intensification separated geographically by the river Ganga down the middle of the state of UP. The high irrigation intensity transitions (middle panel) from *kharif* to multiple cropping occur north and east of the Ganga in the water-rich part of the state whereas the moderate

Table 4.3.1: Summary Statistics

Variable (Unit)	Correlated With	Mean	Std. Deviation
Dependent Variable			
Annual Δpre-Monsoon GWL (mtr)*	Δg	-0.19	1.47
Land-use			
Area HI transition (sq km) ($HI_{i,t}$)**	H_{out}, H_{in}	20.05	17.83
Area MI transition (sq km) ($MI_{i,t}$)**	H_{out}, H_{in}	53.46	52.1
Markets and Infrastructure			
MSP Kharif Crop 1 (INR/100 kg)***	H_{out}, H_{in}	2563	1375
MSP Kharif Crop 2 (INR/100 kg)***	H_{out}, H_{in}	1307	92
MSP Kharif Crop 3 (INR/100 kg)***	H_{out}, H_{in}	2641	1536
MSP Rabi Crop 1 (INR/100 kg)***	H_{out}, H_{in}	3051	50
MSP Rabi Crop 2 (INR/100 kg)***	H_{out}, H_{in}	3026	25
MSP Rabi Crop 3 (INR/100 kg)***	H_{out}, H_{in}	1376	25
Dist. to nearest national hwy (km) $dnh_{i,t}$	H_{out}, H_{in}	10.54	8.11
Climate			
Monsoon precipitation (mtr) ($p_{i,t}$)	$R_{p,i,t}$	0.9	0.08
Summer temp. (K) ($t2m_{i,t}$)	D_{et}, D_{sw}, R_{sw}	307 (33 deg C)	0.4
Geography/Hydro-geology			
Soil Carbon Density (kg m^{-2}) $sc_{i,t}$	H_{out}, H_{in}	6.27	1.15
Dist. to nearest river (km)	D_{sw}, R_{sw}	7.04	4.24
Aquifer type (dummy)	D_{gw}, R_{gw}		
Latitude (dec deg)	All	26.21	0.32
Longitude (dec deg)	All	80.35	0.46
Urban Demand			
Dist. to nearest major city (km) $dub_{i,t}$	H_{out}, H_{in}	56.28	24.52
Total population $pop_{i,t}$	H_{out}, H_{in}	6911	6705
Policy			
CGWB Block class (dummy) $cl_{i,t}$	H_{out}, H_{in}		
Observations	555		

*Dependent variable

**Endogenous variables of interest

***Instruments

	Second Stage
(Intercept)	827.11^{*}
	(338.91)
hi1.sqkm (HI)	-0.03
	(0.03)
mi1.sqkm (MI)	-0.02^{*}
	(0.01)
L0RF (p)	$2.95^{\cdot}$
	(1.62)
L0T2M ($t2m$)	-2.44^{*}
	(0.97)
CAT (cl)	-0.48^{**}
	(0.16)
NEAR_DIST_RIVER (dr)	2.19
	(2.17)
LATITUDE	-3.13^{*}
	(1.37)
LONGITUDE	0.04
	(0.34)
Num. obs.	555

$^{***}p < 0.001$; $^{**}p < 0.01$; $^{*}p < 0.05$; $^{\cdot}p < 0.1$

Wu-Hausman test is significant at $p < 0.1$,

Sargan test is significant at $p < 0.001$,

Wald test: 4.783 on 8 and 546 DF, p-value: 1.064e-05

Table 4.4.1: 2-stage least squares estimation, specification 1 given by equations 4.11.

irrigation transitions from *rabi* (top panel) occur south and west of the Ganga in the relatively drier and hotter parts of the state.

The bottom panel shows annual GWL change aggregated to the sub-districtual level in each year. There is spatial correlation both between HI and MI transitions and GWL depletion (yellow and red areas in the bottom panel).

Rainfall is significant with every additional meter of rainfall causing an increase in GWL change by 2.95 meters. Average summer temperature is also significant with every additional degree Kelvin causing a decrease in GWL change by 2.44 meters. CGWB block categories are significant with semi-critical blocks ($cl_{i,t}$ = 0) having an average change of 0.48 meters less than safe blocks ($cl_{i,t}$ = 1).

Instruments were sufficiently strong and the Wu-Hausman test confirmed at 0.1 significance level that 2SLS was consistent while OLS was not. Our regression model

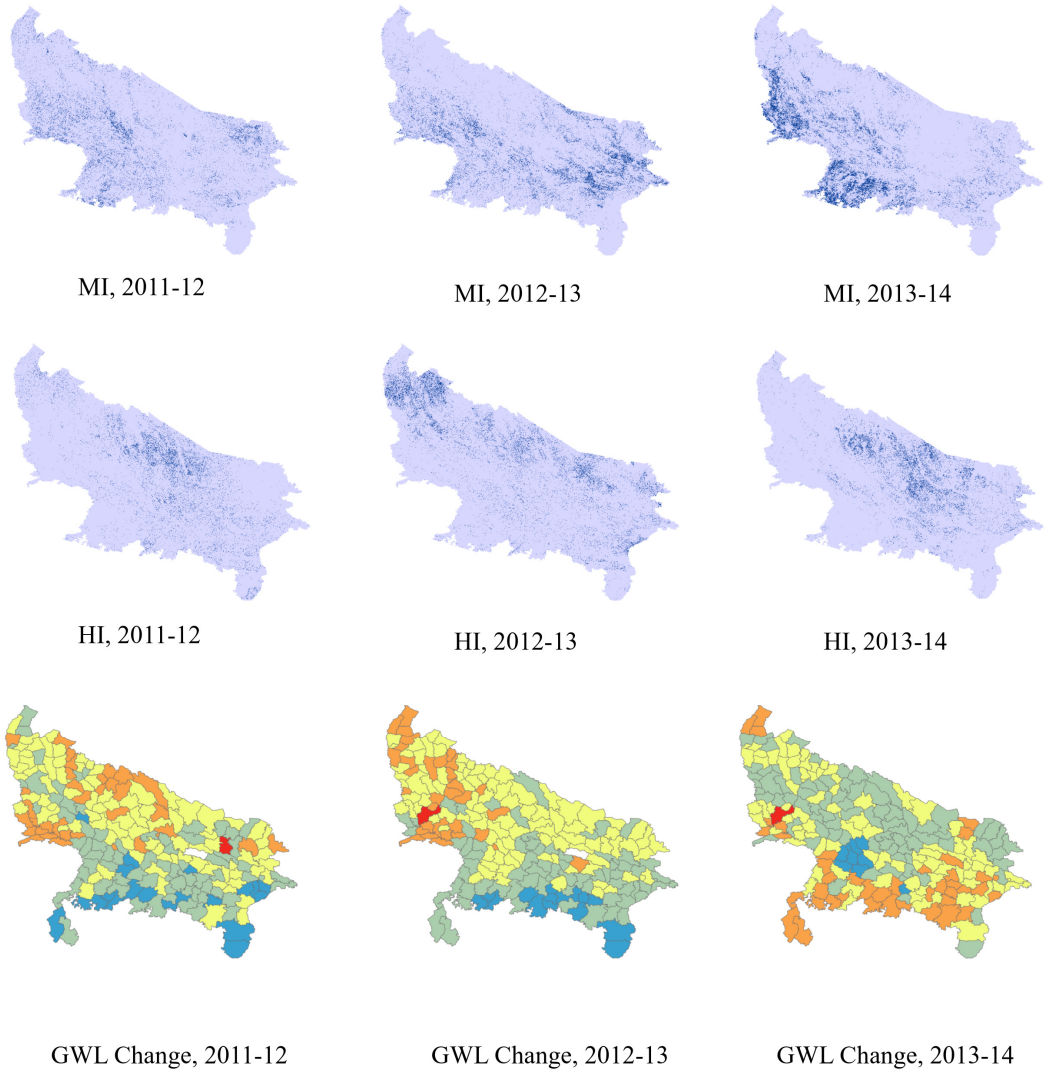


Figure 4.4.1: *Top panel*: Land-use transitions (MI). *Middle panel*: Land-use transitions (HI). Dark blue signals a transition whereas light blue signals none. *Bottom panel*: GWL Change, reds and yellows signal higher depletion. UP, 2011-12 to 2013-14.

Table 4.4.2: First stage regressions

	Dependent variable:	
	mi1.sqkm	hi1.sqkm
	(1)	(2)
CarbonDensity (scd)	-4.893^{**}	1.137
	(2.302)	(0.761)
NEAR_DIST_CITY (dub)	10.061	-18.737^{***}
	(16.806)	(5.555)
NEAR_DIST_NHWY (dnh)	-95.984^{***}	37.587^{***}
	(32.575)	(10.766)
sum_tot_popula (pop)	-0.001^{**}	0.0001
	(0.0004)	(0.0001)
MSP.KC1 (msp^1)	-0.003	0.015
	(0.046)	(0.015)
MSP.KC2 (msp^2)	0.022	0.005
	(0.046)	(0.015)
MSP.KC3 (msp^3)	0.188^{**}	-0.072^{**}
	(0.094)	(0.031)
MSP.RC1 (msp^4)	-0.147	0.158^{***}
	(0.123)	(0.041)
L0RF (p)	-131.455	4.713
	(95.920)	(31.702)
L1T2M ($t2m$)	-8.586	-7.272
	(23.623)	(7.808)
CAT (cl)	-278.151^{***}	88.961^{***}
	(98.108)	(32.425)
NEAR_DIST_RIVER (dr)	54.124	38.645^{*}
	(62.078)	(20.517)
LATITUDE	1.694	-25.039^{**}
	(37.660)	(12.447)
LONGITUDE	-18.064	-6.464
	(17.417)	(5.756)
Constant	4,427.531	3,003.738
	(9,132.078)	(3,018.174)
Observations	555	555
R^2	0.081	0.144
Adjusted R^2	0.058	0.122
Residual Std. Error (df = 540)	50.541	16.704
F Statistic (df = 14; 540)	3.417^{***}	6.502^{***}
Note:	*p<0.1; **p<0.05; ***p<0.01	

loses most of the data due to missing values in the MSP instruments. Many crops like potato, peas and sugarcane that underwent considerable increase in area do not have MSPs. Since MSPs vary by district, if we have a missing value for MSPs in a district, we must discard all of the wells in that district. On account of missing MSPs, our data reduces to only 555 complete observations from a total of over 30000 observations. The first stage regression (Table 4.4.1) drops two instruments corresponding to the MSPs of *rabi* crops 1 and 2 on account of singularities due to the fact they are not linearly independent. This likely occurs because of too many missing values. In order to fill this gap, we must resort to using farm harvest prices (FHPs) for non-MSP crops. We also had to reduce the study period to three years on account data restrictions. With the FHP substitution for MSP, even three years of data consists of over 30000 observations which should be sufficient to get consistent estimates.

FHPs may suffer from endogeneity since they are sensitive to local conditions unlike MSPs. Another possibility is to break up the region into two different sub-regions corresponding to the predominance of MI and HI transitions. The regression, the way it stands, may be confounded by the fact the MSPs of *rabi* crops may not be correlated with increases in MI transitions that occur through *kharif*.

4.5 Discussion

We find evidence of two regimes of land-use transition in Uttar Pradesh from 2011 to 2013. The MI transitions incur a smaller irrigation load and these are concentrated in the region South and West of the Ganga river. These transitions are made by farmers who cropped only during the *rabi* season in the previous year but then went to multiple cropping, presumably by adding a second crop during the *kharif* season, in the present year. The HI transitions that incur heavier irrigation load are made by farmers who practiced *kharif* irrigation in the previous year and went on to add a crop during either *rabi* or *zaid* in the present year thus adding a crop that necessarily relies on irrigation. These transitions occured primarily East of the Ganga river.

The outcome variable in the 2SLS regression is the change in GWL between the current year t and the previous year $t-1$. If GWL this year was deeper (higher in

value), then the change is positive. For example, if GWL this year was 5 m and last year was 4.5 m then the outcome variable is 0.5 m. So when depletion from year to year increases, the change correspondingly increases in magnitude. For example, if GWL in the next year went up to 8 m then the change for the next year would be 8 - 5 = 3 m. We find that an increase in land-use transitions of the MI and HI kind is associated with a reduction in the value of change. This can be interpreted to mean that in areas of greater depletion (higher change in GWL from year to year), fewer land-use transitions occur from single to multiple cropping.

4.6 Conclusion

In this chapter, we started with a structural groundwater balance equation and used it as the basis to develop a framework for the estimation of the causal impact of agricultural land-use intensification on GWL dynamics in the UP region using a 2SLS model. While estimation is still in the preliminary stages due to data restrictions and the fact that we have non-MSP crops in the mix, our coefficients are interpretable and accord with visual evidence. We have elucidated two regimes of land-use intensification through two separate cropping seasons and shown that these are geographically separated by the river Ganga. Preliminary results suggest that fewer land-use transitions to multiple cropping occur in areas where depletion from year to year is greater and that this effect is higher farmers who currently practice *rabi* cropping add a second crop in the *kharif* season.

Appendix

4.A Agricultural Intensification in Uttar Pradesh: A Brief Historical Essay

We present a brief historical account of the evolving patterns of agricultural land-use intensification in UP from the middle Mughal period (1595) onward to the study period of this paper (2009-10 to 2018-2019). The historical account serves as a backdrop against which to interpret the results of the present study. Specifically, it enables us to compare and contrast the rate of present agricultural expansion to the the rates of expansion during selected periods in history. By such comparison, we aim to comprehend whether the current "unsustainable" rates of agricultural intensification and irrigation water withdrawals constitute an unusual departure from what was typically observed in the past.

Mukherji (2022) has argued that the unsustainable irrigation practices of the 'present' emanated from post-Independence agricultural reforms that preceded the "Green Revolution"-era (1967-78 to 1977-1978). Statistical records maintained by Mughal courts aid us in starting our exploration much further back in history than the Green-Revolution era; from the year 1595 during the regnal period of the emperor Akbar. The spatial pattern of agricultural expansion in UP from the year 1595 to the year 1910 in the later British period have been computed using archival data by Moosvi (Moosvi, 2016). Gross cropped area (GCA) in UP *approximately doubled* during this 316 year period. A characteristic feature of this expansion in area was that it followed a west to east trend, closely matching the west to east trends of increasing rainfall and rising (getting closer to the surface) groundwater levels (see Figure 4.A.2). In the water-rich eastern province of Gorakhpur, the GCA increased by a 1000% while progressively smaller increases were recorded moving west toward the relatively drier and hotter imperial capital in Agra where the

increase was a meagre 11% (Moosvi, 2016) (see Figure 4.A.1).

Expansion in irrigation and cropping from 1595 to 1910 appears to have occurred in accord with natural water endowments, a pattern characterized by Shah as "adaptive" irrigation development (Shah, 2010). Whitcombe noted that during the second half of the 19th century in the Mathura[3] province, cropping decisions followed natural conditions in each locality. Rainfed cropping predominated in regions where the groundwater level was far from the surface. Multiple cropping or *do-fasla*[4] was restricted to small tracts where soils could support such practice without the risk of exhaustion (Whitcombe, University of California, Studies, of California. Center for South, & Studies, 1972). The means of irrigation during Mughal times were constructed wells, tanks and canals (Habib, 1999, p.28). While the Mughal administration took some interest in canal construction (Habib, 1999, p.31-34), there is no evidence that innovation in irrigation technology was pursued as a mode of agricultural development. The primary incentive for the Mughal administration to extend the area of cultivation was the land revenue obtained as a result. Expansion in irrigated area, even if groundwater was close to the surface, was constrained by the labor intensive lift mechanisms[5] that were common in Mughal times (Habib, 1999).

These reforms were proposed by international development agencies and enacted by the Indian government with the stated aims of improving agricultural productivity and ensuring food security for independent India (Subramanian, 2015). The reforms were aimed at ensuring national food security and promoted tubewell irrigation with an outlook of realizing the 'full' potential of untapped national groundwater resource. Farmers were subsequently given electricity subsidy and minimum support prices to promote paddy and wheat production (Mukherji, 2022). Mukherji's claim is supported by the fact that in 1966, the western district of Aligarh was the single district in UP to be selected as the site of the new agricultural strategy of intensive cultivation and tubewell powered irrigation. The continued development of groundwater resources that sustained India through a food crisis in the past needs to occur in a "managed" fashion to ensure sustainable futures for India's food production (Shah, 2010; Dantwala, 1976).

[3]This is the modern spelling, the original spelling during the British period was "Muttra."

[4]Also *do-fasli*, literally "double harvest"

[5]The *arhat* or 'Persian wheel' and the *charas*, a bucket that is pulled up by yoked oxen (Habib, 1999, pp.28)

Agricultural land-use change in India in the 20th century was characterized by the expansion of net cultivated area in the first half of the century and the expansion of multiple cropping in the second half of the century. The major expansion in net cropped area in the post-Independence years occurred in the 1950s, the fastest growth occurring between 1941 and 1961 (Kurosaki, 2007). In the 1960s and later years, very little new land was brought into cultivation and expansion of cultivated area was achieved via multiple cropping of existing agricultural lands (Abel, 1970) (see Figure 4.1.1, left panel). According to official land-use statistics, in the part of the GB covered by UP, gross cropped area (GCA) increased by approximately 20% between 1966 and 2000. After the year 2000, official statistics report relatively small (between 1% and 6%) increases in GCA whereas our estimates from high resolution remotely sensed data suggest increases of the order of 50%; a huge deviation from official figures. Irrespective of data source, the rate of agricultural expansion in UP since independence is at least 1.8 times faster than the average rate of expansion in a three hundred year pre-electric period (1595-1910) during the British and Mughal eras (Moosvi, 2016).

There is evidence that groundwater irrigation was a strong driver of agricultural expansion after independence (1950s) up until the 1990s (Dayal, 1977; Narain & Roy, 1980; Dhawan & Datta, 1992) although some economists questioned such findings (Sawant, 1975). Between 1966 and 2012, gross irrigated area (GIA) increased by 150% with growth being particularly fast after economic reforms in the 1990s (DES, 2008; ICRISAT, 2015). The concerns around of groundwater depletion first emerged in the 1990s on account of declining groundwater levels (Chandrakanth & Romm, 1990). A declining trend in groundwater levels can be an indicator of groundwater depletion which is the reduction of water storage beyond natural replenishment usually due to excessive extraction of water for human use. In the Northern Plains of the Indian subcontinent, the euphoria around groundwater development in the years after independence gave way to an urgent need for groundwater resource management in the years leading up to millennium (Shah, 2010; Subramanian, 2015; Dantwala, 1976). In the 2000s, a large and influential research literature has highlighted that the agricultural sector is the largest user of groundwater in India (Anuraga et al., 2006; Rodell et al., 2009; Siebert et al., 2010; Scanlon et al., 2012; Wada et al., 2014; Ahmed & Umar, 2009; B.M.Jha, n.d.). An estimated 89% (217.61 billion cubic meter) of the total withdrawal in 2020 was for

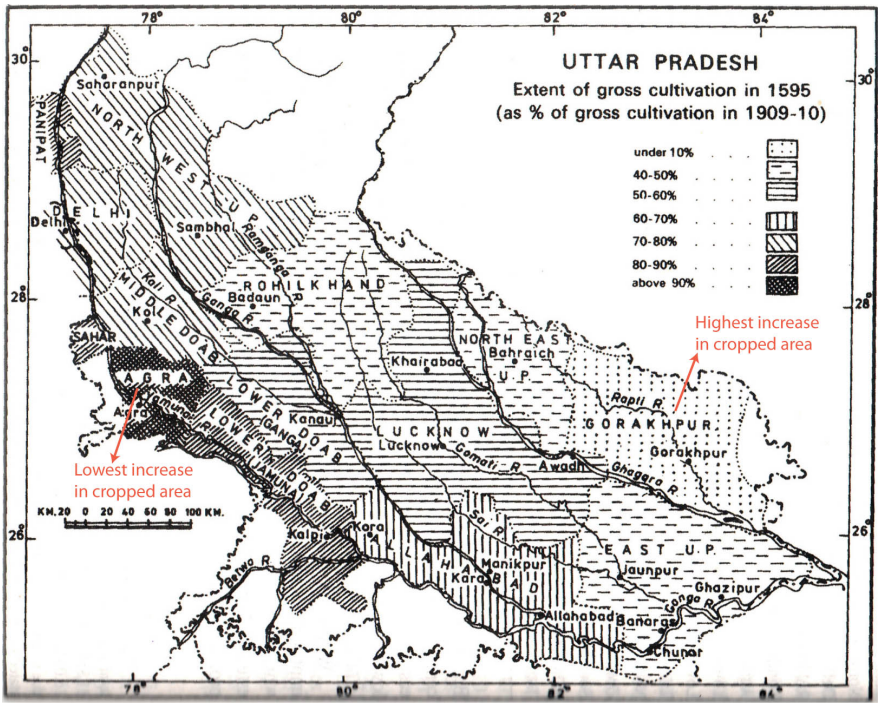


Figure 4.A.1: Change in gross cropped area, 1595-1909 (Moosvi, 2016).

irrigation (CGWB, 2021). The paradigm shift from development to management has occurred on account of severe yet spatially non-uniform groundwater depletion across the sub-continent (CGWB, 2019; Rodell et al., 2009).

The shift from an adaptive to a capitalist mode of agricultural production in the colonial and post-Independence irrigation development has been discussed by Whitcombe (Whitcombe et al., 1972), Attwood (Attwood, 1987), Stone (Stone, 2002), Gilmartin (Gilmartin, 1994), Subramanian (Subramanian, 2015), Dayal (Dayal, 1977) among others (Shah, 2010; Mukherji, 2022). According to a nationwide study by Dayal (Dayal, 1977) spanning the period from 1950 to 1970, the area under multiple cropping in India increased by 50 percent in the 20 year study period (Dayal, 1977). In UP, while multiple cropping increased in the wet northern and eastern parts, they had significantly increased in the relatively drier western and southern parts of the state as well by 1970 (Dayal, 1977) (see Figure 4.A.3).

By contrast, the patterns of agricultural expansion in decades since the "Green Revolution" era (mid-1960s) do not exhibit noticeable accord with the east-west trend in water availability. Areas of heavy agricultural intensification are concentrated in the relatively drier south-central and southwestern parts of UP with pockets of heavy ex-

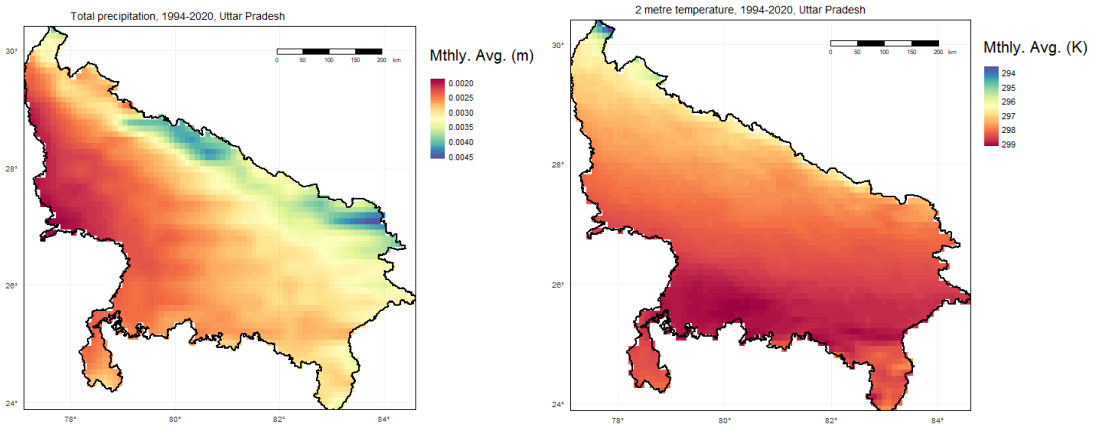


Figure 4.A.2: Spatial Distribution of Rainfall (left) and Temperature (right), 1994-2020, Monthly Means, Source: ERA5 Monthly Averaged Data, Copernicus Climate Data Store (Muñoz Sabater, 2019)

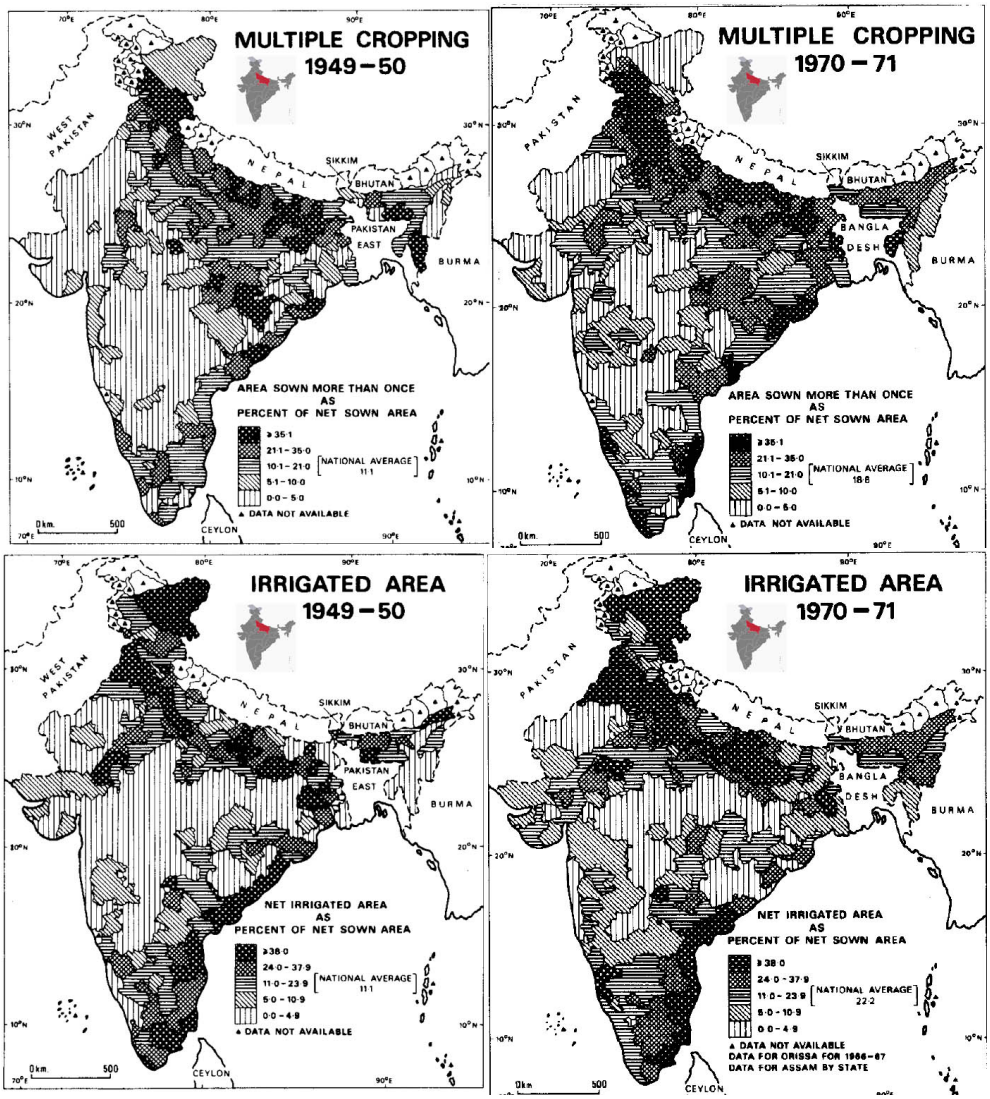


Figure 4.A.3: Agricultural Land-Use Change, 1949-50 and 1970-1961. Percentage multiple cropped (top) and irrigated area (bottom) in 1949-50 and 1970-1961 (Dayal, 1977).

pansion in northern and eastern UP . Groundwater development and over-exploitation is considerably greater in western, southwestern and south-central UP compared to eastern and northern UP (CGWB, 2021).

CHAPTER 5

CONCLUSION

It is important to rigorously characterize and quantify the causal relationship between agricultural and irrigation intensification and groundwater depletion. Such characterization is relevant to groundwater management and public policy and to ensure food security and adapt to a changing climate. The study of groundwater level dynamics and agricultural land-use depends on extensive and expensive groundwater monitoring activities in which data availability and quality can be compromised due to the costs and complexity of monitoring. Further, the spatially dependent nature of groundwater aquifers presents a challenge for econometric models where independence of observations is a necessary condition. The endogeneity of land-use to GWL is an additional methodological challenge.

We rely on missing data theory, geostatistical modeling and a 2SLS regression model to develop a framework in which three important econometric issues can be characterized and partially addressed. We present a rigorous and formal characterization and novel findings on non-random missingness due to dry wells in GWLs data. Our characterization produces an actionable policy recommendation that dry wells must be explicitly recorded and reported by monitoring agencies and the knowledge of dry wells opens the doorway to applying existing methods like censored regression models to improve least squares estimates. Given that monitoring is expensive, the reportage of dry wells can help us improve estimation and inference without any additional infrastructural expenditure.

The spatial nature of the groundwater pumping externality is well studied in the economics and hydrology literature but to quantify the spatial extent of such an externality has not received much attention in economics. A spatial random function model and a functional instrument called a semivariogram aid us to empirically characterize the range of spatial dependence using observed GWLs data. We estimate that GWLs are spatially autocorrelated up to a distance of approximately 14 km in the alluvial aquifers

underlying the Gangetic basin in the state of UP in India. We further propose a strategy to use the semivariogram to calculate the covariance matrix of GWLs and employ a GLS model for consistent estimation. We demonstrate spatial prediction of GWLs using the best linear unbiased estimator called the kriging estimator.

Using a well-level instrumental variables approach, we propose a framework for estimating the causal impact of agricultural land-use intensification on GWL depletion. We employ high resolution spatio-temporal data from remote sensing to elucidate two separate intensification regimes that entail high and moderate use of groundwater irrigation. We also propose the use of a novel instrument; the minimum-support price of crops.

While we have looked at these issues one at a time, in future work we anticipate an integrated model that incorporates strategies for missing data issues, spatial autocorrelation structure and endogeneity. Semivariograms present the possibility of using spatial prediction to impute missing values. They also can enable the use of spatial regression models in which the impact of spatially proximate GWLs can be explicitly incorporated in a regression.

This work advances the study of groundwater level dynamics and agricultural land-use intensification bringing statistical theory to the service of groundwater management and demonstrates results in the context of the Gangetic Basin covering parts of the important food producing state of UP in North India. While rooted in the specific study region, the econometric strategies proposed in this work generalize to all groundwater management and policy making contexts.

References

Abbasi, A., DiTraglia, F. J., Gazze, L., & Pals, B. (2023). Hidden hazards and screening policy: Predicting undetected lead exposure in illinois. *Journal of Health Economics*, 102783.

Abel, M. E. (1970). Agriculture in india in the 1970s. *Economic and Political Weekly*, *5*(13), A5–A14. Retrieved from http://www.jstor.org/stable/4359788

Agricultural Census. (2016). *Average size of operational land holdings, uttar pradesh.* Directorate of Economics and Statistics, Department of Agriculture and Famer Welfare, Goverment of India. Retrieved from https://www.ceicdata.com/en/india/agriculture-census-average-size-of-operational-land-holdings-by-size-group/agriculture-census-average-size-of-operational-land-holdings-uttar-pradesh-size-group-all-holdings (Online @CEIC. Retrieved June 23, 2023)

Ahmed, I., & Umar, R. (2009, 10). Groundwater flow modelling of yamuna-krishni interstream, a part of central ganga plain uttar pradesh. *Journal of Earth System Science*, *118*, 507-523. doi: 10.1007/s12040-009-0050-5

Ali, S., & Arora, G. (2021, August). *Well-level Missingness Mechanisms in Administrative Groundwater Monitoring Data for Uttar Pradesh (UP), India, 2009-2018* (2021 Annual Meeting, August 1-3, Austin, Texas). Agricultural and Applied Economics Association. Retrieved from https://ageconsearch.umn.edu/record/314038/files/Abstracts_21_06_30_14_11_54_69__106_210_96_147_0.pdf doi: 10.22004/ag.econ.314038

Ali, S., & Arora, G. (2022). Spatial auto-correlation in groundwater levels. In *2022 annual meeting, july 31-august 2, anaheim, california.*

Ali, S., Liu, D., Fu, Q., Cheema, M. J. M., Pham, Q. B., Rahaman, M. M., ... Anh, D. T.

(2021). Improving the resolution of grace data for spatio-temporal groundwater storage assessment. *Remote Sensing*, *13*(17). Retrieved from https://www.mdpi.com/2072-4292/13/17/3513 doi: 10.3390/rs13173513

Alley, W. (2009). Ground water. In G. E. Likens (Ed.), *Encyclopedia of inland waters* (p. 684-690). Oxford: Academic Press. Retrieved from https://www.sciencedirect.com/science/article/pii/B9780123706263000156 doi: https://doi.org/10.1016/B978-012370626-3.00015-6

Altonji, J. G., Bharadwaj, P., & Lange, F. (2012). Changes in the characteristics of american youth: Implications for adult outcomes. *Journal of Labor Economics*, *30*(4), 783-828. Retrieved from https://doi.org/10.1086/666536 doi: 10.1086/666536

Amemiya, T. (1984). Tobit models: A survey. *Journal of econometrics*, *24*(1-2), 3–61.

Anselin, L. (1988). Spatial heterogeneity. In *Spatial econometrics: Methods and models* (pp. 119–136). Dordrecht: Springer Netherlands. Retrieved from https://doi.org/10.1007/978-94-015-7799-1_9 doi: 10.1007/978-94-015-7799-1_9

Anselin, L. (2013). *Spatial econometrics: Methods and models*. Springer Netherlands. Retrieved from https://books.google.co.in/books?id=G47tCAAAQBAJ

Anselin, L., & Bera, A. K. (1998). Spatial dependence in linear regression models with an introduction to spatial econometrics. *Statistics textbooks and monographs*, *155*, 237–290.

Anselin, L., et al. (2001). Spatial econometrics. *A companion to theoretical econometrics*, *310330*.

Anuraga, T., Ruiz, L., Kumar, M. M., Sekhar, M., & Leijnse, A. (2006). Estimating groundwater recharge using land use and soil data: A case study in south india. *Agricultural Water Management*, *84*(1), 65 - 76. Retrieved from http://www.sciencedirect.com/science/article/pii/S0378377406000412 doi: https://doi.org/10.1016/j.agwat.2006.01.017

Arora, G., & Ali, S. (2022). Climate resilience in the presence of water-related stress

and hazards in india. *EAERE Magazine*.

Arora, G., Rathore, T., Gupta, G., & Anand, S. (2019). Socioeconomic and biophysical drivers of cropland use intensification in india: Analysis using satellite data and administrative surveys.

Asoka, A., Gleeson, T., Wada, Y., & Mishra, V. (2017). Relative contribution of monsoon precipitation and pumping to changes in groundwater storage in india. *Nature Geoscience*, *10*(2), 109–117.

Asoka, A., & Mishra, V. (2019). Groundwater pumping to increase food production causes persistent groundwater drought in india. *arXiv preprint arXiv:1908.00255*.

Attwood, D. W. (1987). Irrigation and imperialism: The causes and consequences of a shift from subsistence to cash cropping. *The Journal of Development Studies*, *23*(3), 341-366. Retrieved from https://doi.org/10.1080/00220388708422037 doi: 10.1080/00220388708422037

Bárdossy, A. (1997). Introduction to geostatistics. *Institute of Hydraulic Engineering, University of Stuttgart*.

Beale, C. M., Lennon, J. J., Yearsley, J. M., Brewer, M. J., & Elston, D. A. (2010). Regression analysis of spatial data. *Ecology letters*, *13*(2), 246–264.

BGS. (2022a). *Groundwater level information.* British Geological Survey. Retrieved from https://www2.bgs.ac.uk/groundwater/datainfo/levels/home.html (Retrieved on September 8th, 2022)

BGS. (2022b). *Groundwater level terminology.* British Geological Survey. Retrieved from https://www2.bgs.ac.uk/groundwater/datainfo/levels/terminology.html#:~:text=Groundwater%20level%20is%20a%20term,top%20of%20the%20saturated%20zone. (Retrieved on November 27th, 2022)

Bhanja, S. N., Mukherjee, A., Rangarajan, R., Scanlon, B. R., Malakar, P., & Verma, S. (2019). Long-term groundwater recharge rates across india by in situ measurements. *Hydrology and Earth System Sciences*, *23*(2), 711–722.

Bhanja, S. N., Mukherjee, A., Rodell, M., Wada, Y., Chattopadhyay, S., Velicogna, I., ... Famiglietti, J. S. (2017). Groundwater rejuvenation in parts of india influenced by water-policy change implementation. *Scientific reports*, *7*(1), 1–7.

Bhanja, S. N., Mukherjee, A., Saha, D., Velicogna, I., & Famiglietti, J. S. (2016). Validation of grace based groundwater storage anomaly using in-situ groundwater level measurements in india. *Journal of Hydrology*, *543*, 729 - 738. Retrieved from http://www.sciencedirect.com/science/article/pii/S0022169416306886 doi: https://doi.org/10.1016/j.jhydrol.2016.10.042

Bhattarai, N., Lobell, D. B., null, Fishman, R., Kustas, W. P., Pokhrel, Y., & Jain, M. (2023). Warming temperatures exacerbate groundwater depletion rates in india. *Science Advances*, *9*(35), eadi1401. Retrieved from https://www.science.org/doi/abs/10.1126/sciadv.adi1401 doi: 10.1126/sciadv.adi1401

Bhattarai, N., Pollack, A., Lobell, D. B., Fishman, R., Singh, B., Dar, A., & Jain, M. (2021a). The impact of groundwater depletion on agricultural production in india. *Environmental Research Letters*, *16*(8), 085003.

Bhattarai, N., Pollack, A., Lobell, D. B., Fishman, R., Singh, B., Dar, A., & Jain, M. (2021b, jul). The impact of groundwater depletion on agricultural production in india. *Environmental Research Letters*, *16*(8), 085003. Retrieved from https://doi.org/10.1088/1748-9326/ac10de doi: 10.1088/1748-9326/ac10de

Biemans, H., Siderius, C., Mishra, A., & Ahmad, B. (2016). Crop-specific seasonal estimates of irrigation-water

demand in south asia. *Hydrology and Earth System Sciences*, *20*(5), 1971–1982. Retrieved from https://hess.copernicus.org/articles/20/1971/2016/ doi: 10.5194/hess-20-1971-2016

Bivand, R. S., Pebesma, E. J., Gomez-Rubio, V., & Pebesma, E. J. (2013). *Applied spatial data analysis with r* (Vol. 2). Springer.

B.M.Jha. (n.d.). Towards better management of ground water resources in india. Retrieved from http://cgwb.gov.in/documents/papers/incidpapers/Paper%201-B.M.Jha.pdf

Bonsor, H., Macdonald, A., Ahmed, K. M., Burgess, W., Basharat, M., Calow, R., ... Zahid, A. (2017, 02). Hydrogeological typologies of the indo-gangetic basin alluvial aquifer, south asia. *Hydrogeology Journal*, *25*, 1377-1406. doi: 10.1007/

s10040-017-1550-z

Bresciani, E., Shandilya, R. N., Kang, P. K., & Lee, S. (2020). Well radius of influence and radius of investigation: What exactly are they and how to estimate them? *Journal of Hydrology*, *583*, 124646. Retrieved from https://www.sciencedirect.com/science/article/pii/S0022169420301062 doi: https://doi.org/10.1016/j.jhydrol.2020.124646

Briggs, A., Clark, T., Wolstenholme, J., & Clarke, P. (2003, May). Missing.... presumed at random: cost-analysis of incomplete data. *Health Economics*, *12*(5), 377-392. Retrieved from https://ideas.repec.org/a/wly/hlthec/v12y2003i5p377-392.html doi: 10.1002/hec.766

Brown, S. J., Goetzmann, W., Ibbotson, R. G., & Ross, S. A. (1992). Survivorship bias in performance studies. *The Review of Financial Studies*, *5*(4), 553–580.

Brozović, N., Sunding, D. L., & Zilberman, D. (2010). On the spatial nature of the groundwater pumping externality. *Resource and Energy Economics*, *32*(2), 154–164.

Burkhauser, R. V., Butler, J., & Kim, Y. W. (1995). The importance of employer accommodation on the job duration of workers with disabilities: A hazard model approach. *Labour Economics*, *2*(2), 109–130.

CACP. (2022, December). *Determinants of MSP*. Commission for Agricultural Costs and Prices, Ministry of Agriculture and Farmers Welfare, Government of India. Retrieved from https://cacp.dacnet.nic.in/content.aspx?pid=62 (Retrieved on December 14th, 2022)

Census of India. (2011). *Census of India.* Office of the Registrar General Census Commissioner, India. Retrieved from https://censusindia.gov.in/census.website/ (Retrieved on August 28th, 2022)

CGWB. (n.d.). *Aquifer mapping and groundwater management plan, parts of ncr, uttar pradesh.* Ministry of Water Resources, River Development and Ganga Rejuvenation, Government of India. Retrieved from http://cgwb.gov.in/AQM/NAQUIM_REPORT/UP/NCR,%20U.P.pdf

CGWB. (2009). *Report of the Groundwater Resource Estimation Committee.* Ministry of Water Resources, Government of India. Retrieved from http://cgwb.gov.in/Documents/GEC97.pdf (Retrieved on October 10th, 2022)

CGWB. (2015). *Concept note on national project on aquifer management.* Government of India. Retrieved from http://cgwb.gov.in/

CGWB. (2017). *Detailed Guidelines for Implementing the Ground Water Resource Estimation Methodology 2015.* Ministry of Water Resources, Government of India. Retrieved from http://cgwb.gov.in/GW-Assessment/2020-11-17_Detailed_Guidelines_GEC2015.pdf (Retrieved on November 2nd, 2022)

CGWB. (2017, June). *Dynamic Groundwater Resource of India (As on 31st March 2013).* Ministry of Water Resources, River Development & Ganga Rejuvenation, Government of India, Faridabad. Retrieved from http://cgwb.gov.in/Documents/Dynamic%20GWRE-2013.pdf (Retrieved on November 2nd, 2022)

CGWB. (2017). *Report of the Groundwater Resource Estimation Committee (GEC-2015).* Ministry of Water Resources, River Development & Ganga Rejuvenation Government of India. Retrieved from http://cgwb.gov.in/Documents/GEC2015_Report_Final%2030.10.2017.pdf (Retrieved on November 1st, 2022)

CGWB. (2018). *District-wise Depth to Water Table, Premonsoon, 2018.* Department of Water Resources, River Development & Ganga Rejuvenation, Ministry of Jal Shakti, Government of India. Retrieved from http://cgwb.gov.in/documents/Uttar%20Pradesh.pdf (Retrieved on September 25th, 2022)

CGWB. (2019a). *Central Groundwater Board FAQ.* Retrieved from http://cgwb.gov.in/faq.html

CGWB. (2019b). *Central Groundwater Board Website.* Retrieved from http://cgwb.gov.in/aboutcgwb.html

CGWB. (2019c). *Ground Water Year Book, Uttar Pradesh, 2019 -2020.* Department of Water Resources, River Development & Ganga Rejuvenation, Ministry of Jal Shakti, Government of India. Retrieved from URL:http://cgwb.gov.in/Regions/NR/Reports/Year%20Book%2019-20.pdf (Retrieved on May 22nd 2022)

CGWB. (2019, July). *National Compilation on Dynamic Groundwater Re-*

sources of India 2017. Department of Water Resources, River Development & Ganga Rejuvenation, Ministry of Jal Shakti, Government of India. Retrieved from http://cgwb.gov.in/GW-Assessment/GWRA-2017-National-Compilation.pdf (Retrieved on November 5th, 2020.)

CGWB. (2021, December). *Ground Water Year Book, Uttar Pradesh, 2020 -2021*. Department of Water Resources, River Development & Ganga Rejuvenation, Ministry of Jal Shakti, Government of India. Retrieved from http://cgwb.gov.in/Regions/NR/Reports/Year%20Book_CGWB_NR%202020-21.pdf (Retrieved June 6th 2022)

CGWB. (2021, July). *National Compilation on Dynamic Groundwater Resources of India 2020*. Department of Water Resources, River Development & Ganga Rejuvenation, Ministry of Jal Shakti, Government of India. Retrieved from http://cgwb.gov.in/documents/2021-08-02-GWRA_India_2020.pdf (Retrieved on July 21st, 2022.)

Chandrakanth, M., & Romm, J. (1990). Groundwater depletion in india—institutional management regimes. *Natural Resources Journal*, *30*(3), 485–501. Retrieved 2022-07-10, from http://www.jstor.org/stable/24883588

Chung, S. Y., Senapathi, V., .S, S., & Prasanna, M. (2019, 06). Supplement of missing data in groundwater-level variations of peak type using geostatistical methods.. doi: 10.1016/B978-0-12-815413-7.00004-3

C Koreimann, G. W. W. N., J Grath, & Vogel, W. R. (1996). *Groundwater monitoring in europe* (Tech. Rep.). Retrieved from https://www.eea.europa.eu/publications/92-9167-032-4/download (Retrieved 08-09-2022)

Cobb, A. R., & Harvey, C. F. (2019). Scalar simulation and parameterization of water table dynamics in tropical peatlands. *Water Resources Research*, *55*(11), 9351–9377.

Cressie, N. (2015). *Statistics for spatial data*. John Wiley & Sons.

Cressie, N., Cressie, A., & Sons, J. W. . (1991). *Statistics for spatial data*. J. Wiley. Retrieved from https://books.google.co.in/books?id=k_JQAAAAMAAJ

Cressie, N., & Moores, M. T. (2021). Spatial statistics. *arXiv preprint arXiv:2105.07216*.

Cressie, N. A. (1993). *Statistics for spatial data* (Tech. Rep.).

Cunningham, W., & Schalk, C. (2014). *Groundwater technical procedures of the u.s. geological survey*. Createspace Independent Pub. Retrieved from https://books.google.co.in/books?id=TgaMoAEACAAJ

Cázares Escareño, J., Júnez-Ferreira, H. E., González-Trinidad, J., Bautista-Capetillo, C., & Robles Rovelo, C. O. (2022). Design of groundwater level monitoring networks for maximum data acquisition at minimum travel cost. *Water*, *14*(8). Retrieved from https://www.mdpi.com/2073-4441/14/8/1209 doi: 10.3390/w14081209

Dantwala, M. (1976). *Agricultural policy in india since independence* (1976 Conference, July 26-August 4, 1976, Nairobi, Kenya No. 182350). International Association of Agricultural Economists. Retrieved from https://EconPapers.repec.org/RePEc:ags:iaae76:182350

Das, J., Rahman, A. S., Mandal, T., & Saha, P. (2020). Challenges of sustainable groundwater management for large scale irrigation under changing climate in lower ganga river basin in india. *Groundwater for Sustainable Development*, *11*, 100449.

Das, S. (2011). *Groundwater resources of india*. National Book Trust, India. Retrieved from https://books.google.co.in/books?id=y7BMMwEACAAJ

Dax, A. (1985). Completing missing groundwater observations by interpolation. *Journal of Hydrology*, *81*(3-4), 375–399.

Dax, A., & Zilberbrand, M. (2017, 07). Imputing missing groundwater observations. *Hydrology Research*, *49*(3), 831-845. Retrieved from https://doi.org/10.2166/nh.2017.220 doi: 10.2166/nh.2017.220

Dayal, E. (1977). Impact of irrigation expansion on multiple cropping in india*. *Tijdschrift voor Economische en Sociale Geografie*, *68*(2), 100-109. Retrieved from https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1467-9663.1977.tb01399.x doi: https://doi.org/10.1111/j.1467-9663.1977.tb01399.x

Delhomme, J. (1978). Kriging in the hydrosciences. *Advances in Water Resources*, *1*(5), 251-266. Retrieved from https://www.sciencedirect.com/science/article/pii/0309170878900398 doi: https://doi

.org/10.1016/0309-1708(78)90039-8

DES. (2008). *District-wise Land Use Statistics, Directorate of Economics and Statistics, Government of India.* Ministry of Agriculture, Government of India. Retrieved from https://aps.dac.gov.in/LUS/Public/Reports.aspx ([Online. Retrieved 24-09-2021])

Dhawan, B. D., & Datta, H. S. (1992). Impact of irrigation on multiple cropping. *Economic and Political Weekly*, *27*(13), A15–A18. Retrieved 2022-07-21, from http://www.jstor.org/stable/4397728

Dickinson, R. (2014, September 16). *AB-1739.* Retrieved from https://leginfo.legislature.ca.gov/faces/billStatusClient.xhtml?bill_id=201320140AB1739 (Retrieved on November 1st, 2022)

Dong, Y., Jiang, C., Suri, M. R., Pee, D., Meng, L., & Rosenberg Goldstein, R. E. (2019). Groundwater level changes with a focus on agricultural areas in the mid-atlantic region of the united states, 2002–2016. *Environmental Research*, *171*, 193-203. Retrieved from https://www.sciencedirect.com/science/article/pii/S0013935119300039 doi: https://doi.org/10.1016/j.envres.2019.01.004

Doran, H., & Griffiths, W. (1983). On the relative efficiency of estimators which include the initial observations in the estimation of seemingly unrelated regressions with first-order autoregressive disturbances. *Journal of Econometrics*, *23*(2), 165-191. Retrieved from https://www.sciencedirect.com/science/article/pii/030440769390075G doi: https://doi.org/10.1016/0304-4076(93)90075-G

Edwards, E. C. (2016). What lies beneath? aquifer heterogeneity and the economics of groundwater management. *Journal of the Association of Environmental and Resource Economists*, *3*(2), 453-491. Retrieved from https://doi.org/10.1086/685389 doi: 10.1086/685389

Ekpe, G. K. (2021). *Pumping fees and spillovers in the groundwater commons: An evaluation of a conservation tool and irrigator competitive behavior* (Doctoral dissertation). Retrieved from https://www.proquest.com/openview/d3eb12fd279ba924c9fd9d218d222fff/1?pq-origsite=

gscholar&cbl=18750&diss=y

Enders, C. K. (2010). *Applied missing data analysis*. Guilford press.

Evans, S., Williams, G. P., Jones, N. L., Ames, D. P., & Nelson, E. J. (2020). Exploiting earth observation data to impute groundwater level measurements with an extreme learning machine. *Remote Sensing*, *12*(12). Retrieved from https://www.mdpi.com/2072-4292/12/12/2044 doi: 10.3390/rs12122044

Famiglietti, J. S. (2014). The global groundwater crisis. *Nature Climate Change*, *4*(11), 945–948.

Farsi aliabadi, M. M., Daneshvar Kakhki, M., Sabouhi, M., Dourandish, A., & Amadeh, H. (2020, 10). Effect of water conservation policies on groundwater depletion in iran. *Journal of Chinese Soil and Water Conservation*, *51*, 109-116. doi: 10.29417/JCSWC.202009_51(3).0003

F. Dormann, C., M. McPherson, J., B. Araújo, M., Bivand, R., Bolliger, J., Carl, G., ... others (2007). Methods to account for spatial autocorrelation in the analysis of species distributional data: a review. *Ecography*, *30*(5), 609–628.

Fishman, R. (2018, March). Groundwater depletion limits the scope for adaptation to increased rainfall variability in India. *Climatic Change*, *147*(1), 195-209. doi: 10.1007/s10584-018-2146-x

Fletcher, T., Andrieu, H., & Hamel, P. (2013). Understanding, management and modelling of urban hydrology and its consequences for receiving waters: A state of the art. *Advances in Water Resources*, *51*, 261-279. Retrieved from https://www.sciencedirect.com/science/article/pii/S0309170812002412 (35th Year Anniversary Issue) doi: https://doi.org/10.1016/j.advwatres.2012.09.001

Forder, J., & Allan, S. (2013, 12). The impact of competition on quality and prices in the english care homes market. *Journal of health economics*, *34C*, 73-83. doi: 10.1016/j.jhealeco.2013.11.010

Gardner, R., Ostrom, E., & Walker, J. M. (1990). The nature of common-pool resource problems. *Rationality and society*, *2*(3), 335–358.

Gilmartin, D. (1994). Scientific empire and imperial science: Colonialism and irrigation technology in the indus basin. *The Journal of Asian Studies*, *53*(4), 1127–1149. Retrieved from http://www.jstor.org/stable/2059236

Gisser, M., & Sanchez, D. A. (1980). Competition versus optimal control in groundwater pumping. *Water resources research*, *16*(4), 638–642.

Goovaerts, P. (1997, 01). Geostatistics for natural resource evaluation. In (Vol. 42).

Government of Uttar Pradesh. (n.d.). *Uttar Pradesh Groundwater Department (UPGWD)*. Retrieved from http://upgwd.gov.in/StaticPages/Background-hi.aspx (Retrieved on December 4th, 2020)

Graham, J. W. (2009). Missing data analysis: Making it work in the real world. *Annual Review of Psychology*, *60*(1), 549-576. Retrieved from https://doi.org/10.1146/annurev.psych.58.110405.085530 (PMID: 18652544) doi: 10.1146/annurev.psych.58.110405.085530

Griffith, D. A. (2019). Negative spatial autocorrelation: One of the most neglected concepts in spatial statistics. *Stats*, *2*(3), 388–415. Retrieved from https://www.mdpi.com/2571-905X/2/3/27 doi: 10.3390/stats2030027

Guhathakurta, P., L, S. K. B., Menon, P., Prasad, A. K., Sable, S. T., & Advani, S. C. (2020). *Observed rainfall variability and changes overuttar pradesh state.* Indian Meteorological Department. Retrieved from https://imdpune.gov.in/hydrology/rainfall%20variability%20page/uttar_final.pdf ([Online. Retrieved 20-09-2021])

Gulati, A., Terway, P., & Hussain, S. (2021, 03). Performance of agriculture in uttar pradesh. In (p. 175-210). doi: 10.1007/978-981-15-9335-2_7

Gundogdu, K. S., & Guney, I. (2007). Spatial analyses of groundwater levels using universal kriging. *Journal of earth system science*, *116*(1), 49–55.

Gupta, D. (2021, February). *Free Power, Irrigation and Groundwater Depletion: Impact of the Farm Electricity Policy of Punjab, India* (Working papers No. 316). Centre for Development Economics, Delhi School of Economics. Retrieved from https://ideas.repec.org/p/cde/cdewps/316.html

Guzmán, S. M., Paz, J. O., Tagert, M. L. M., Mercer, A. E., & Pote, J. W. (2018). An integrated svr and crop model to estimate the impacts of irrigation on daily groundwater levels. *Agricultural systems*, *159*, 248–259.

Habib, I. (1999). *The agrarian system of mughal india, 1556-1707*. Oxford University Press. Retrieved from https://books.google.co.in/

Hasan, K., Paul, S., Chy, T. J., & Antipova, A. (2021, July). Analysis of groundwater table variability and trend using ordinary kriging: the case study of Sylhet, Bangladesh. *Applied Water Science*, *11*(7), 120. doi: 10.1007/s13201-021-01454-w

Heiss, F. (2011). Dynamics of self-rated health and selective mortality. *Empirical economics*, *40*, 119–140.

Hora, T., Srinivasan, V., & Basu, N. (2019, 08). The groundwater recovery paradox in south india. *Geophysical Research Letters*, *46*. doi: 10.1029/2019GL083525

Hrozencik, R. A., Manning, D. T., Suter, J. F., Goemans, C., & Bailey, R. T. (2017). The heterogeneous impacts of groundwater management policies in the republican river basin of colorado. *Water Resources Research*, *53*, 10,757 - 10,778.

Huang, J., Cao, L., Yu, F., Liu, X., & Wang, L. (2021). Groundwater drought and cycles in xuchang city, china. *Frontiers in Earth Science*, *9*. Retrieved from https://www.frontiersin.org/article/10.3389/feart.2021.736305 doi: 10.3389/feart.2021.736305

Hussain, F., Wu, R.-S., & Shih, D.-S. (2022). Water table response to rainfall and groundwater simulation using physics-based numerical model: Wash123d. *Journal of Hydrology: Regional Studies*, *39*, 100988.

ICRISAT. (2015). *Updated district level-database (apportioned, 1966 district boundaries)*. Author. Retrieved from http://data.icrisat.org/dld/src/about-dld.html ([Online. Retrieved 24-09-2021])

India Observatory. (2022). *India Observatory Data Platform.* Retrieved from https://dp.observatory.org.in/content/about-data-platform (Retrieved on September 25st, 2022)

India-WRIS. (2022). *India Water Resource Information System.* Government of India. Retrieved from http://www.india-wris.nrsc.gov.in/ (https://indiawris.gov.in/wris/)

IRMED. (n.d.). Institutional framework for regulating use of ground water in india. Retrieved from http://cgwb.gov.in/incgw/kamta%20prasad%20report.pdf (Submitted by Institute for Resource Management and Economic Development. Sponsored by Ministry of Water Resources, Government of India.)

ISEC. (2003, December). *Impact of Minimum Support Prices on Agricultural Economy.* Agricultural Development and Rural Transformation Unit, Institute for Social and Economic Change. Retrieved from https://cacp.dacnet.nic.in/ViewContents.aspx?Input=1&PageId=36&KeyId=0 (Retrieved on December 14th, 2022)

Islam, Z., Ranganathan, M., Bagyaraj, M., Singh, S. K., & Gautam, S. K. (2021). Multi-decadal groundwater variability analysis using geostatistical method for groundwater sustainability. *Environment, Development and Sustainability*, 1–19.

ISRO. (2022). *Bhuvan: Indian Geo-Platform of ISRO.* National Remote Sensing Center, ISRO. Retrieved from https://bhuvan.nrsc.gov.in/bhuvan_links.php

Jain, M., Fishman, R., Mondal, P., Galford, G. L., Bhattarai, N., Naeem, S., ... others (2021). Groundwater depletion will reduce cropping intensity in india. *Science Advances*, *7*(9), eabd2849.

Jamali Jaghdani, T. (2012). *Demand for irrigation water from depleting groundwater resources: An econometric approach* (Unpublished doctoral dissertation).

Jha, K., & McKinley, C. (2014, 07). Demography and ecology of indian sarus crane (grus antigone antigone) in uttar pradesh, northern india. *Asian Journal of Conservation Biology*, *3*.

Johnson, A. (1963). *Compilation of specific yield for various materials* (Tech. Rep.).

Jukić, D., & Denić-Jukić, V. (2009). Groundwater balance estimation in karst by using a conceptual rainfall–runoff model. *Journal of hydrology*, *373*(3-4), 302–315.

Kareemulla, K. (2022). *International and national agricultural research system in india.* National Academy of Agricultural Research Management. Retrieved from https://naarm.org.in/focarsrepository/files/1.%20International%20and%20National%20Agricultural%20Research%20System%20in%20India.pdf (Retrieved on November 20th, 2022)

Kenda, K., Koprivec, F., & Mladenić, D. (2018). Optimal missing value estimation algorithm for groundwater levels. *Multidisciplinary Digital Publishing Institute Proceedings*, *2*(11), 698.

Khan, H. H., & Khan, A. (2019). Chapter 14 - groundwater and surface water interac-

tion. In S. Venkatramanan, M. V. Prasanna, & S. Y. Chung (Eds.), *Gis and geostatistical techniques for groundwater science* (p. 197-207). Elsevier. Retrieved from https://www.sciencedirect.com/science/article/pii/B9780128154137000146 doi: https://doi.org/10.1016/B978-0-12-815413-7.00014-6

King, G., & Zeng, L. (2001). Logistic regression in rare events data. *Political Analysis*, *9*(2), 137–163. doi: 10.1093/oxfordjournals.pan.a004868

Kirby, J., Ahmad, M., Mainuddin, M., Palash, W., Quadir, M., Shah-Newaz, S., & Hossain, M. (2015). The impact of irrigation development on regional groundwater resources in bangladesh. *Agricultural Water Management*, *159*, 264–276.

Konikow, L. F., & Kendy, E. (2005). Groundwater depletion: A global problem. *Hydrogeology Journal*, *13*(1), 317–320.

Krige, D. G. (1951). A statistical approach to some basic mine valuation problems on the witwatersrand, by d.g. krige, published in the journal, december 1951 : introduction by the author. *Journal of The South African Institute of Mining and Metallurgy*, *52*, 201-203.

Kurosaki, T. (2007, 12). Land-use changes and agricultural growth in india, pakistan, and bangladesh, 1901-2004. *Land-Change Science in the Tropics: Changing Agricultural Landscapes*. doi: 10.1007/978-0-387-78864-7_4

Lapworth, D. J., Dochartaigh, B. , Nair, T., O'Keeffe, J., Krishan, G., MacDonald, A. M., ... Jackson, C. R. (2021). Characterising groundwater-surface water connectivity in the lower Gandak catchment, a barrage regulated biodiversity hotspot in the mid-Gangetic basin. *Journal of Hydrology*, *594*, 125923. Retrieved from https://www.sciencedirect.com/science/article/pii/S0022169420313846 doi: https://doi.org/10.1016/j.jhydrol.2020.125923

Lauridsen, J. (2012). Spatial economics. In S. J. Smith (Ed.), *International encyclopedia of housing and home* (p. 1-4). San Diego: Elsevier. Retrieved from https://www.sciencedirect.com/science/article/pii/B9780080471631006834 doi: https://doi.org/10.1016/B978-0-08-047163-1.00683-4

Legendre, P. (1993). Spatial autocorrelation: trouble or new paradigm? *Ecology*, *74*(6), 1659–1673.

Li, H., & Zhao, J. (2018). Rebound effects of new irrigation technologies: The role of water rights. *American Journal of Agricultural Economics*, *100*, 786–808.

Lins, H. F., & Cohn, T. A. (2011). Stationarity: wanted dead or alive? 1. *JAWRA Journal of the American Water Resources Association*, *47*(3), 475–480.

Little, R. J. (1988). A test of missing completely at random for multivariate data with missing values. *Journal of the American statistical Association*, *83*(404), 1198–1202.

Little, R. J., & Rubin, D. B. (2019). *Statistical analysis with missing data* (Vol. 793). John Wiley & Sons.

Liu, H.-L., Bao, A., Pan, X.-L., & Chen, X. (2013, 08). Effect of land-use change and artificial recharge on the groundwater in an arid inland river basin. *Water Resources Management*, *27*. doi: 10.1007/s11269-013-0380-6

Lockwood, D. (2021). *Fooled by the winners: How survivor bias deceives us*. Greenleaf Book Group. Retrieved from https://books.google.co.in/books?id=xTBAEAAAQBAJ

Long, D., Chen, X., Scanlon, B., Wada, Y., Hong, Y., Singh, V., ... Yang, W. (2016, 04). Have grace satellites overestimated groundwater depletion in the northwest india aquifer? *Scientific Reports*, *6*. doi: 10.1038/srep24398

MacDonald, A., Bonsor, H., Ahmed, K., Burgess, W., Basharat, M., Calow, R., ... others (2016). Groundwater quality and depletion in the indo-gangetic basin mapped from in situ observations. *Nature Geoscience*, *9*(10), 762–766.

Machiwal, D., Mishra, A., Jha, M. K., Sharma, A., & Sisodia, S. (2012). Modeling short-term spatial and temporal variability of groundwater level using geostatistics and gis. *Natural resources research*, *21*(1), 117–136.

Maréchal, J., Galeazzi, L., & Dewandel, B. (2003). Importance ofirrigation return flow on the groundwater budget of a rural basin in india. In *Hydrology of mediterranean and semiarid regions: Papers selected for the international conference on hydrology of the mediterranean and semi-arid regions, held in montpellier, france from 1 to 4 april 2003* (p. 62).

Matheron, G. (1963). Principles of geostatistics. *Economic geology*, *58*(8), 1246–1266.

Matheron, G. (1971). *The theory of regionalized variables and its applications*. École national supérieure des mines. Retrieved from https://books.google.co

.in/books?id=TGhGAAAAYAAJ

Mealli, F., & Rubin, D. B. (2015). Clarifying missing at random and related definitions, and implications when coupled with exchangeability. *Biometrika*, *102*(4), 995–1000.

Meghwal, R., Shah, D., & Mishra, V. (2019). On the changes in groundwater storage variability in western india using grace and well observations. *Remote Sensing in Earth Systems Sciences*, *2*(4), 260–272.

Merrill, N. H., & Guilfoos, T. (2018). Optimal groundwater extraction under uncertainty and a spatial stock externality. *American journal of agricultural economics*, *100*(1), 220–238.

MI Census. (2013a). *Fifth census of minor irrigation schemes.* Ministry of Jal Shakti, Department of Water Resources, River Development & Ganga Rejuvenation, Ministry of Jal Shakti, Government of India. Retrieved from http://164.100.229.38/sites/default/files/5thMISchemes-Eng_0.pdf (Retrieved on October 31, 2022)

MI Census. (2013b). *Minor irrigation census.* Ministry of Jal Shakti, Department of Water Resources, River Development & Ganga Rejuvenation, Ministry of Jal Shakti, Government of India. Retrieved from http://164.100.229.38/state-wise-reports?shs_term_node_tid_depth_1=88&shs_term_node_tid_depth=125 (Retrieved on September 25th, 2022)

Milly, P., Betancourt, J., Falkenmark, M., Hirsch, R. M., Kundzewicz, Z. W., Lettenmaier, D. P., & Stouffer, R. J. (2008). Stationarity is dead: Whither water management? *Earth*, *4*, 20.

Ministry of Jal Shakti. (2020, March). *Atal Bhujal Yojana Program Guidelines, Version 1.1.* Department of Water Resources, River Development & Ganga Rejuvenation, Government of India. Retrieved from http://jalshakti-dowr.gov.in/sites/default/files/Atal_Bhujal_Yojana_Program_Guidelines_Ver_1.pdf (Retrieved on October 21st, 2022)

Moosvi, S. (2016). *The economy of the mughal empire, c.1595: A statistical study*. Oxford University Press. Retrieved from https://books.google.co.in/

Moran, P. A. P. (1950, 06). Notes on Continuous Stochastic Phenomena. *Biometrika*, *37*(1-2), 17-23. Retrieved from https://doi.org/10.1093/biomet/37.1-2.17 doi: 10.1093/biomet/37.1-2.17

Moukana, J. A., & Koike, K. (2008). Geostatistical model for correlating declining groundwater levels with changes in land cover detected from analyses of satellite images. *Computers & Geosciences*, *34*(11), 1527-1540. Retrieved from https://www.sciencedirect.com/science/article/pii/S0098300408000393 doi: https://doi.org/10.1016/j.cageo.2007.11.005

Mukherjee, A., Bhanja, S. N., & Wada, Y. (2018). Groundwater depletion causing reduction of baseflow triggering ganges river summer drying. *Scientific reports*, *8*(1), 12049.

Mukherjee, A., & Ramachandran, P. (2018). Prediction of gwl with the help of grace tws for unevenly spaced time series data in india : Analysis of comparative performances of svr, ann and lrm. *Journal of Hydrology*, *558*, 647-658. Retrieved from https://www.sciencedirect.com/science/article/pii/S0022169418300817 doi: https://doi.org/10.1016/j.jhydrol.2018.02.005

Mukherjee, S., Mishra, A., & Trenberth, K. E. (2018). Climate change and drought: a perspective on drought indices. *Current Climate Change Reports*, *4*(2), 145–163.

Mukherji, A. (2022, March). Sustainable Groundwater Management in India Needs a Water-Energy-Food Nexus Approach. *Applied Economic Perspectives and Policy*, *44*(1), 394-410. Retrieved from https://ideas.repec.org/a/wly/apecpp/v44y2022i1p394-410.html doi: 10.1002/aepp.13123

Muñoz Sabater, J. (2019). *ERA5-Land monthly averaged data from 1981 to present.* Copernicus Climate Change Service (C3S) Climate Data Store (CDS). (Retrieved on December 9th, 2020) doi: 10.24381/cds.68d2bb30

Nageswara Rao, G., Babu, P., & Bantilan, C. (2009, 01). Dynamics and development pathways in the semi-arid tropics: Dokur village profile. research bulletin no. 23.

Narain, D., & Roy, S. (1980). *Impact of irrigation and labor availability on multiple cropping: a case study of India* (Tech. Rep.).

Narayanamoorthy, A. (2015). Groundwater depletion and water extraction cost: some evidence from south india. *International Journal of Water Resources Development*, *31*(4), 604-617. Retrieved from https://doi.org/10.1080/

07900627.2014.935302 doi: 10.1080/07900627.2014.935302

NRSC. (2007). *National Land Use and Land Cover Mapping Using Multi-temporal AWiFS data.* National Remote Sensing Center, Department of Space, Government of India. Retrieved from https://bhuvan-app1.nrsc.gov.in/2dresources/thematic/LULC250/0506.pdf (Retrieved on March 5th, 2021)

Ostrom, E. (1990). *Governing the commons: The evolution of institutions for collective action*. Cambridge university press.

Panda, D. K., & Wahr, J. (2016). Spatiotemporal evolution of water storage changes in i ndia from the updated grace-derived gravity records. *Water Resources Research*, *52*(1), 135–149.

Pandey, G., Khan, R., & Kumari, S. (2019). Pathways of agricultural transformation: a comparative analysis of punjab and bihar. *Agricultural Economics Research Review*, *32*(347-2020-1010).

Pant, N. (1994). The turnover of public tube wells in uttar pradesh: A case study of a successful cooperative society. In *International conference on irrigation management transfer* (pp. 20–24).

Parasyris, A., Spanoudaki, K., Varouchakis, E. A., & Kampanis, N. A. (2021, 07). A decision support tool for optimising groundwater-level monitoring networks using an adaptive genetic algorithm. *Journal of Hydroinformatics*, *23*(5), 1066-1082. Retrieved from https://doi.org/10.2166/hydro.2021.045 doi: 10.2166/hydro.2021.045

Pavley, F. (2014a, September 16). *SB-1168.* Retrieved from https://leginfo.legislature.ca.gov/faces/billStatusClient.xhtml?bill_id=201320140AB1739 (Retrieved on November 1st, 2022)

Pavley, F. (2014b, September 16). *SB-1319.* Retrieved from https://leginfo.legislature.ca.gov/faces/billNavClient.xhtml?bill_id=201320140SB1319 (Retrieved on November 1st, 2022)

Pebesma, E. J. (2004). Multivariable geostatistics in s: the gstat package. *Computers & Geosciences*, *30*(7), 683–691.

Perrone, D., & Jasechko, S. (2017a). Dry groundwater wells in the western united

states. *Environmental Research Letters*, *12*(10), 104002.

Perrone, D., & Jasechko, S. (2017b, sep). Dry groundwater wells in the western united states. *Environmental Research Letters*, *12*(10), 104002. Retrieved from https://dx.doi.org/10.1088/1748-9326/aa8ac0 doi: 10.1088/1748-9326/aa8ac0

Pfeiffer, L., & Lin, C.-Y. C. (2012). Groundwater pumping and spatial externalities in agriculture. *Journal of Environmental Economics and Management*, *64*(1), 16–30.

Pfeiffer, L., & Lin, C.-Y. C. (2014). Does efficient irrigation technology lead to reduced groundwater extraction? empirical evidence. *Journal of Environmental Economics and Management*, *67*, 189-208.

Qiu, J., et al. (2010). China faces up to groundwater crisis. *Nature*, *466*(7304), 308–308.

Rodell, M., Velicogna, I., & Famiglietti, J. S. (2009). Satellite-based estimates of groundwater depletion in India. *Nature*, *460*(7258), 999.

Rubin, D. B. (1976). Inference and missing data. *Biometrika*, *63*(3), 581–592.

Salas-Eljatib, C., Fuentes-Ramirez, A., Gregoire, T. G., Altamirano, A., & Yaitul, V. (2018). A study on the effects of unbalanced data when fitting logistic regression models in ecology. *Ecological Indicators*, *85*, 502–508.

Sarkar, T., Kannaujiya, S., Taloor, A. K., Ray, P. K. C., & Chauhan, P. (2020). Integrated study of grace data derived interannual groundwater storage variability over water stressed indian regions. *Groundwater for Sustainable Development*, *10*, 100376.

Sawant, S. (1975). Extent of multiple cropping in irrigated and unirrigated areas of india: Some implications for usefulness of irrigation statistics. *Indian Journal of Agricultural Economics*, *30*(902-2018-1108), 48–53.

Sayre, S., & Taraz, V. (2018, 11). Groundwater depletion in india: Social losses from costly well deepening. *Journal of Environmental Economics and Management*. doi: 10.1016/j.jeem.2018.11.002

Scanlon, B. R., Faunt, C. C., Longuevergne, L., Reedy, R. C., Alley, W. M., McGuire, V. L., & McMahon, P. B. (2012). Groundwater depletion and sustainability of irrigation in the us high plains and central valley. *Proceedings of the national*

academy of sciences, *109*(24), 9320–9325.

Schoengold, K., & Brozovic, N. (2018). The future of groundwater management in the high plains: evolving institutions, aquifers and regulations. *Western Economics Forum*, *16*(1). Retrieved from https://ideas.repec.org/a/ags/weecfo/273678.html doi: 10.22004/ag.econ.273678

Scibek, J., & Allen, D. (2006). Modeled impacts of predicted climate change on recharge and groundwater levels. *Water Resources Research*, *42*(11).

Seah, K. K. (2018). Immigrant educators and students' academic achievement. *Labour Economics*, *51*, 152–169.

Sears, L., Lim, D., & Lawell, C.-Y. C. L. (2018). The economics of agricultural groundwater management institutions: The case of california. *Water Economics and Policy*.

Sekhri, S. (2014). Wells, water, and welfare: the impact of access to groundwater on rural poverty and conflict. *American Economic Journal: Applied Economics*, *6*(3), 76–102.

Semiromi, M. T., & Koch, M. (2019). Reconstruction of groundwater levels to impute missing values using singular and multichannel spectrum analysis: application to the ardabil plain, iran. *Hydrological Sciences Journal*, *64*(14), 1711-1726. Retrieved from https://doi.org/10.1080/02626667.2019.1669793 doi: 10.1080/02626667.2019.1669793

Shah, T. (2010). *Taming the anarchy: Groundwater governance in south asia*. Taylor & Francis. Retrieved from https://books.google.co.in/books?id=cF_sp7UcuGoC

Shortridge, A. (n.d.). *Course notes, ordinary kriging.* Retrieved from https://msu.edu/~ashton/classes/866/papers/gatrell_ordkrige.pdf ([Online; Retrieved 24-12-2020])

Siebert, S., Burke, J., Faures, J. M., Frenken, K., Hoogeveen, J., Döll, P., & Portmann, F. T. (2010). Groundwater use for irrigation - a global inventory. *Hydrology and Earth System Sciences*, *14*(10), 1863–1880. Retrieved from https://www.hydrol-earth-syst-sci.net/14/1863/2010/ doi: 10.5194/hess-14-1863-2010

Singh, N., Kudrat, M., Jain, K., & Pandey, K. (2011, 08). Cropping pattern of uttar

pradesh using irs-p6 (awifs) data. *International Journal of Remote Sensing*, *32*, 4511-4526. doi: 10.1080/01431161.2010.489061

Smith, S. M. (2018). Economic incentives and conservation: Crowding-in social norms in a groundwater commons. *Journal of Environmental Economics and Management*, *90*, 147–174.

Sterk, G. (2021). A hillslope version of the revised morgan, morgan and finney water erosion model. *International Soil and Water Conservation Research*, *9*(3), 319-332. Retrieved from https://www.sciencedirect.com/science/article/pii/S2095633921000149 doi: https://doi.org/10.1016/j.iswcr.2021.01.004

Stone, I. (2002). *Canal irrigation in british india: Perspectives on technological change in a peasant economy*. Cambridge University Press. Retrieved from https://books.google.co.in/books?id=7WLUWxIcyogC

Stralberg, D., & Bao, S. (1999, 12). Identifying the spatial structure in error terms with spatial covariance models: A case study on urbanization influence in chaparral bird species. *Annals of GIS*, *5*, 106-120. doi: 10.1080/10824009909480520

Subramanian, K. (2015). *Revisiting the green revolution: Irrigation and food production in twentieth-century india*. King's College London. Retrieved from https://books.google.co.in/books?id=yM4TvwEACAAJ

Tang, N., & Ju, Y. (2018). Statistical inference for nonignorable missing-data problems: a selective review. *Statistical Theory and Related Fields*, *2*(2), 105-133. Retrieved from https://doi.org/10.1080/24754269.2018.1522481 doi: 10.1080/24754269.2018.1522481

Tobin, J. (1958). Estimation of relationships for limited dependent variables. *Econometrica: journal of the Econometric Society*, 24–36.

Tobler, W. (1970). A computer movie simulating urban growth in the detroit region. *Economic Geography*, *46*, 234-240.

Tobler, W. (1999). Linear pycnophylactic reallocation comment on a paper by d. martin. *Int. J. Geogr. Inf. Sci.*, *13*, 85-90.

Todd, D., & Mays, L. (2005). *Groundwater hydrology*.

Tsur, Y. (1990, 05). The stabilization role of groundwater when surface water supplies are uncertain: The implications for groundwater development. *Wa-*

ter Resources Research - WATER RESOUR RES, *26*, 811-818. doi: 10.1029/WR026i005p00811

Unfried, K., Kis-Katos, K., & Poser, T. (2022). Water scarcity and social conflict. *Journal of Environmental Economics and Management*, *113*, 102633. Retrieved from https://www.sciencedirect.com/science/article/pii/S0095069622000171 doi: https://doi.org/10.1016/j.jeem.2022.102633

Unwin, D. (2009). Statistics, spatial. In R. Kitchin & N. Thrift (Eds.), *International encyclopedia of human geography* (p. 452-457). Oxford: Elsevier. Retrieved from https://www.sciencedirect.com/science/article/pii/B9780080449104005393 doi: https://doi.org/10.1016/B978-008044910-4.00539-3

UPGWD. (2019). *Uttar Pradesh Groundwater (Management and Regulation) Act 2019.* Uttar Pradesh Ground Water Department. Retrieved from http://upgwdonline.in/images/GWDact2019En.pdf (Retrieved on November 2nd, 2022)

UPGWD. (2020). *Pre & Post Monsoon Ground Water Level Data.* Uttar Pradesh Ground Water Department. Retrieved from http://upgwd.gov.in/MediaGallery/WLD08092017.pdf (Retrieved November 5th, 2020)

USGS. (1902). *Water-supply and irrigation papers of the united states geological survey* (No. v. 67). U.S. Government Printing Office. Retrieved from https://books.google.co.in/books?id=L3JMAAAAYAAJ

USGS. (2003). *Ground-water depletion across the nation, usgs fact sheet 103-03.* ([Online; accessed 31-December-2020, https://pubs.usgs.gov/fs/fs-103-03/])

USGS. (2013a). Oregon Water Science Center. Retrieved from https://or.water.usgs.gov/projs_dir/willgw/glossary.html (Glossary of Hydrologic Terms [Online. Retrieved 23-2-2021])

USGS. (2013b). *A National Framework for Ground Water Monitoring in the United States.* The Subcommittee on Ground Water of the Advisory Committee on Water Information. Retrieved from https://cida.usgs.gov/ngwmn/doc/ngwmn_framework_report_july2013.pdf (Retrieved on September 8th, 2022)

USGS. (2018a). *Groundwater decline and depletion.* US Geological Survey. Retrieved

from https://www.usgs.gov/special-topics/water-science-school/science/groundwater-decline-and-depletion (Retrieved November 27th, 2022)

USGS. (2018b). *Groundwater decline and depletion.* Retrieved from https://www.usgs.gov/special-topics/water-science-school/science/groundwater-decline-and-depletion (Retrieved on June 1st, 2022)

USGS. (2019). *What makes a groundwater well go dry?* Retrieved from https://www.usgs.gov/special-topics/water-science-school/science/water-qa-what-makes-groundwater-well-go-dry (Retrieved on June 1st, 2022)

Varghese, S. K., Veettil, P. C., Speelman, S., Buysse, J., & Van Huylenbroeck, G. (2013). Estimating the causal effect of water scarcity on the groundwater use efficiency of rice farming in south india. *Ecological Economics*, *86*, 55-64. Retrieved from https://www.sciencedirect.com/science/article/pii/S0921800912004041 (Sustainable Urbanisation: A resilient future) doi: https://doi.org/10.1016/j.ecolecon.2012.10.005

Varouchakis, E., Hristopulos, D., & Karatzas, G. (2012). Improving kriging of groundwater level data using nonlinear normalizing transformations—a field application. *Hydrological Sciences Journal*, *57*(7), 1404-1419. doi: 10.1080/02626667.2012.717174

Varouchakis, E. A., Hristopulos, D. T., & Karatzas, G. P. (2009, April). A Study of the Groundwater Level Spatial Variability in the Messara Valley of Crete. In *Egu general assembly conference abstracts* (p. 9351).

Vrigazova, B. (2021). The proportion for splitting data into training and test set for the bootstrap in classification problems. *Business Systems Research: International Journal of the Society for Advancing Innovation and Research in Economy*, *12*(1), 228–242.

Wada, Y., Wisser, D., & Bierkens, M. F. (2014). Global modeling of withdrawal, allocation and consumptive use of surface water and groundwater resources. *Earth System Dynamics Discussions*, *5*(1), 15–40.

Weiskel, P. K., Vogel, R. M., Steeves, P. A., Zarriello, P. J., DeSimone, L. A., &

Ries III, K. G. (2007). Water use regimes: Characterizing direct human interaction with hydrologic systems. *Water Resources Research*, *43*(4).

Whitcombe, E., University of California, B. C. f. S., Studies, S. A., of California. Center for South, U., & Studies, S. A. (1972). *Agrarian conditions in northern india: The united provinces under british rule, 1860-1900*. University of California Press. Retrieved from https://books.google.co.in/books?id=De7ZAAAAMAAJ

Wooldridge, J. (2013). *Introductory Econometrics: A Modern Approach*. Cengage Learning. Retrieved from https://books.google.co.in/books?id=C0KHwUKxys0C

Xiong, J., Abhishek, Guo, S., & Kinouchi, T. (2022). Leveraging machine learning methods to quantify 50 years of dwindling groundwater in india. *Science of The Total Environment*, *835*, 155474. Retrieved from https://www.sciencedirect.com/science/article/pii/S0048969722025700 doi: https://doi.org/10.1016/j.scitotenv.2022.155474

Yan, J., Chen, X., Luo, G., & Guo, Q. (2006). Temporal and spatial variability response of groundwater level to land use/land cover change in oases of arid areas. *Chinese Science Bulletin*, *51*(1), 51–59.

Zaveri, E., Grogan, D., Fisher-Vanden, K., Frolking, S., Lammers, R., Wrenn, D., ... Nicholas, R. (2016, 08). Invisible water, visible impact: Groundwater use and indian agriculture under climate change. *Environmental Research Letters*, *11*, 084005. doi: 10.1088/1748-9326/11/8/084005

Zirulia, A., Barbagli, A., & Guastaldi, E. (2021). Groundwater level forecasting for water resource management. *Applied Modeling Techniques and Data Analysis 1: Computational Data Analysis Methods and Tools*, *7*, 221–231.

LIST OF PAPERS BASED ON THESIS

5.1 Publications

5.1.1 Papers in Refereed Journals/Magazines

1. Climate resilience in the presence of water-related stress and hazards in India
   Gaurav Arora, Saif Ali
   EAERE Magazine, **n17**, 42-49 (Summer-Fall, 2022). Link to paper.

5.1.2 Presentations in Conferences

1. Accepted titled *Spatial Autocorrelation in Groundwater Levels* at the **Research Scholars Day,** Department of Economic Sciences, IIT Kanpur, Kanpur, India, Jan, 2024.

2. Accepted titled *Groundwater level dynamics and agricultural land-use change in the Gangetic Basin* at the **Roorkee Water Conclave,** IIT Roorkee and NIH Roorkee, Roorkee, Uttarakhand, India, Mar 3-6, 2024.

3. Accepted titled *Spatial Autocorrelation in Groundwater Levels* at the **Roorkee Water Conclave,** IIT Roorkee and NIH Roorkee, Roorkee, Uttarakhand, India, Mar 3-6, 2024.

4. Presented titled *Spatial Autocorrelation in Groundwater Levels* at the **Annual Meeting of the Agricultural & Applied Economics Association (AAEA),** Anaheim, CA, Jul 31- Aug 2, 2022. Link to poster.

5. Presented titled *Well Dryness as a Missingness Mechanism in Administrative Groundwater Level Data* at the **Asian Meeting of the Econometric Society (AMES) 2022,** Shenzhen, China, Jun 22 - Jun 24, 2022.

6. Presented titled *Well Dryness as a Missingness Mechanism in Administrative Groundwater Level Data* at the **97th Annual Conference of the Western Economics Association International (WEAI),** Portland, Oregon, Jun 29- Jul 3, 2022.

7. Presented titled *Well Dryness as a Missingness Mechanism in Administrative Groundwater Level Data* at the **48th Annual Conference of the Eastern Economics Association (EEA),** Montego Bay, Jamaica, May 5-7, 2022.

8. Presented titled *Spatial Autocorrelation in Groundwater Levels* at the **48th Annual Conference of the Eastern Economics Association (EEA),** Montego Bay, Jamaica, May 5-7, 2022.

9. Presented titled *Missingness Mechanisms and Spatial Autocorrelation Patterns in Groundwater Levels* at the **Cutting-Edge Issues in Spatial Econometrics and Image Processing: Missing Data, Causal Inference and Machine Learning, DST-JSPS Joint Seminar between Nagasaki University and IIITD,** Nagasaki, Japan and New Delhi, India, Mar 02-03, 2022.

10. Presented titled *Well Dryness as a Missingness Mechanism in Administrative Groundwater Level Data* at the **Winter School, Delhi School of Economics and the Econometric Society, 2021,** Dec 15-18, 2021.

11. Presented titled *Well-level Missingness Mechanisms in Administrative Groundwater Monitoring Data for Uttar Pradesh (UP), India, 2009-2018* at the **2021 Annual Meeting, Agricultural and Applied Economics Association,** Austin, TX, August 1-3, 2021. Link to paper.

12. Presented titled *Spatial Autocorrelation in Groundwater Levels* at the **INSEE Biennial Conference 2021,** New Delhi, India, Dec 15-17, 2021.

13. Presented titled *Cropping intensification, aquifer hydrogeology and groundwater depletion in Western Uttar Pradesh* at the **Indo-US bilateral symposium on "The study of decadal scale droughts and mega-droughts in semi-arid tracts of India and North America",** IISER Mohali, India, January 2, 2020.

14. Presented titled *Sustainable groundwater utilization amidst agricultural intensification?, An assessment of cropping decisions and groundwater levels in the Indo-Gangetic Plains* at the **ASABE/ISAE Global Water Security Conference,** Hyderabad, India, 2018.

15. Presented titled *Sustainable groundwater extraction amidst agricultural land use change and policy., Analysis using satellite imagery (and other) data* at the **Workshop on Computational Social Science,** IIITD, New Delhi, India, 2018.