That Sinking Feeling: Predicting Land Subsidence in California's San Joaquin Valley with a Spatial Regression Model

by

Cole Ira Heap

A Thesis Presented to the
FACULTY OF THE USC DORNSIFE COLLEGE OF LETTERS, ARTS AND SCIENCES
University of Southern California
In Partial Fulfillment of the
Requirements for the Degree
MASTER OF SCIENCE
(GEOGRAPHIC INFORMATION SCIENCE AND TECHNOLOGY)

December 2022

For my sons, Chace, Henry, Connor, and Jax.
To show them that education is a lifelong journey

Acknowledgements

I am grateful for this opportunity to combine my background in geology with spatial analytics. The M.S. in GIST program helped to make this possible. I appreciate all assistance provided to me by USC SSI staff. Their experience and constructive commentary have greatly assisted me in becoming a better GIS professional and spatial modeler. I am also grateful for my peers in the GIST program – you know who you are! Your encouragement, willingness to listen, and ability to bring me back down to earth are truly appreciated! Finally, I would truly miss the mark if it weren't for my wonderful wife, Chong. As she does, she has given me this opportunity to further my education and do something that is very time consuming and taxing on the family. She is the real heroine of this thesis!

Table of Contents

List of Tables

List of Equations

List of Figures

Abbreviations

BGS	Below ground surface
DWR	Department of Water Resources
GAMA	Groundwater Ambient Monitoring and Assessment Program
GIS	Geographic information system
GSA	Groundwater Sustainability Agency
GSP	Groundwater sustainability plan
GTWR	Geographic temporally weighted regression
GWR	Geographically weighted regression
MAUP	Modifiable areal unit problem
RWQCB	Regional Water Quality Control Board
SEM	Spatial error model
SGMA	Sustainable Groundwater Management Act
SLM	Spatial lag model
USGS	United States Geological Survey
VIF	Variance inflation factor

Abstract

Land subsidence is an ongoing problem in California's San Joaquin Valley. Due to drought and over extraction of groundwater, land subsidence occurs at a rate of more than one foot per year. Since California enacted the Sustainable Groundwater Management Act in 2014, land subsidence has been labelled one of the six undesirable effects that causes degradation of groundwater aquifers. Spatially assessing and identifying issues pertinent to land subsidence tends to come after subsidence has already occurred. Modeling land subsidence has been attempted with some success but doing so has required complex hydrogeologic models that are computationally intensive and require large volumes of data to be collected for processing and input. This research incorporated simple, but key geological and engineering variables that are derived from the United States Geological Survey and the California Department of Water Resources. From these sources, a robust dataset was used to statistically explore spatial patterns and relationships among groundwater levels, amount of fine-grained sediment present in the aquifer, confined or unconfined aquifer designation, well completion length, aquitard clay thickness, and well depth all as they pertain to land subsidence. Land subsidence patterns were assessed with exploratory techniques of generalized linear regression and geographically weighted regression. Each method was used to visualize the spatial distribution and scale of land subsidence relationships among groundwater wells from 2015 to 2021. Due to the size of the valley, the number of wells found throughout, and accompanying variability in independent variables, global scale predictions of land subsidence were not as successful as local regression techniques. Geographically weighted regression took into consideration the variance among variables, accounted for spatial autocorrelation, and yielded an easy-to-update, but accurate prediction for spatial patterns of land subsidence in the San Joaquin Valley.

Chapter 1 Introduction

In California, land subsidence has been well documented since the start of the 20th century (Poland 1972). California's subsidence is largely the result of excessive groundwater extraction that when combined with the drier seasons due to climate change has become a rampant challenge in the San Joaquin Valley. Subsidence is at historically high rates of more than one foot per year and poses a risk not only to agriculture but to public and private water supply wells. Quick but accurate forecasting of land subsidence is crucial for effective groundwater management and environmental planning. In this study, spatial patterns of land subsidence in the San Joaquin Valley are assessed using key geological variables.

This study uses recorded land subsidence from 2015 to 2021 to assess spatial relationships among explanatory variables such as annual change in groundwater level, groundwater level, subbasin area, well total depth, well completion length, confining clay layer depth, percentage of fine versus coarse-grained sediment, well vintage, classification of confined vs. unconfined aquifer, and confining clay layer thickness as each are key geological factors that relate to land subsidence.

This study estimates the spatial variation of land subsidence from key geological and hydrological variables. This effort will assist with generating simplified, long-term subsidence estimates comparable to the complex Central Valley Hydrologic Model (CVHM) along with the level of accuracy needed for Groundwater Sustainability Plans (GSPs).

This study evaluates spatial regression models to assess spatial patterns of land subsidence. To create a simple, but accurate regression model independent geologic variables linked to the causation and prediction of land subsidence are used. Ali et al. (2020) and Chu et al. (2021) utilized geographical temporal weighted regression (GTWR), which is an extension of

geographically weighed regression (GWR), which accounts for local effects in space and time. Furthermore, GTWR accounts for local effects of groundwater drawdown and subsidence in space and time making it sufficient for small slices of aggregated time, in this case, on an annual time slice for forecasting land subsidence.

1.1. Land Subsidence and Groundwater

The unique thing about groundwater, outside of it being a relatively abundant resource, is that it has the tendency to keep grains of rock and sediment apart from other grains, ultimately inflating the ground elevation (Fetter 2001). The extraction of water (and other fluids) from the subsurface eliminates the holding of such pore spaces open. Since air is compressible and water is not, this results in the rock grains and sediment compacting. Collapsing such pore space results in the ground above sinking. This is known as land subsidence (Neunedorf et al. 2011).

Figure 1 exhibits land subsidence between 1965 and 2017 near Merced, CA (USGS 2016). Each year notation shows the ground elevation for the designated year. Photos were taken just south of Merced in December 2017 and show how land has dropped 8.6 feet between 1965 and 2016. The rate of subsidence was even greater from 1988 to 2016 when 6.2 feet of elevation loss occurred all due to groundwater extraction. Subsidence, particularly in the San Joaquin Valley, continues today at close to historically high rates of more than one foot per year (USGS 2021).

Figure 1. Measured land subsidence through time outside of Merced, CA

1.2. Study Area

The San Joaquin Valley is a Mediterranean environment with long, hot, and dry summers and wet winters. It lies north of the Transverse Range and south of the Sacramento Valley. The San Joaquin Valley produces close to 13% of the United States' agricultural products which include grapes, raisins, cotton, almonds, citrus, and a plethora of vegetables (CDFA 2010). Each of these crops are water intensive, so pumping groundwater in the Central Valley to water crops is not uncommon. Figure 2 shows the approximate location of the San Joaquin Valley, the area of interest (AOI) for this study.

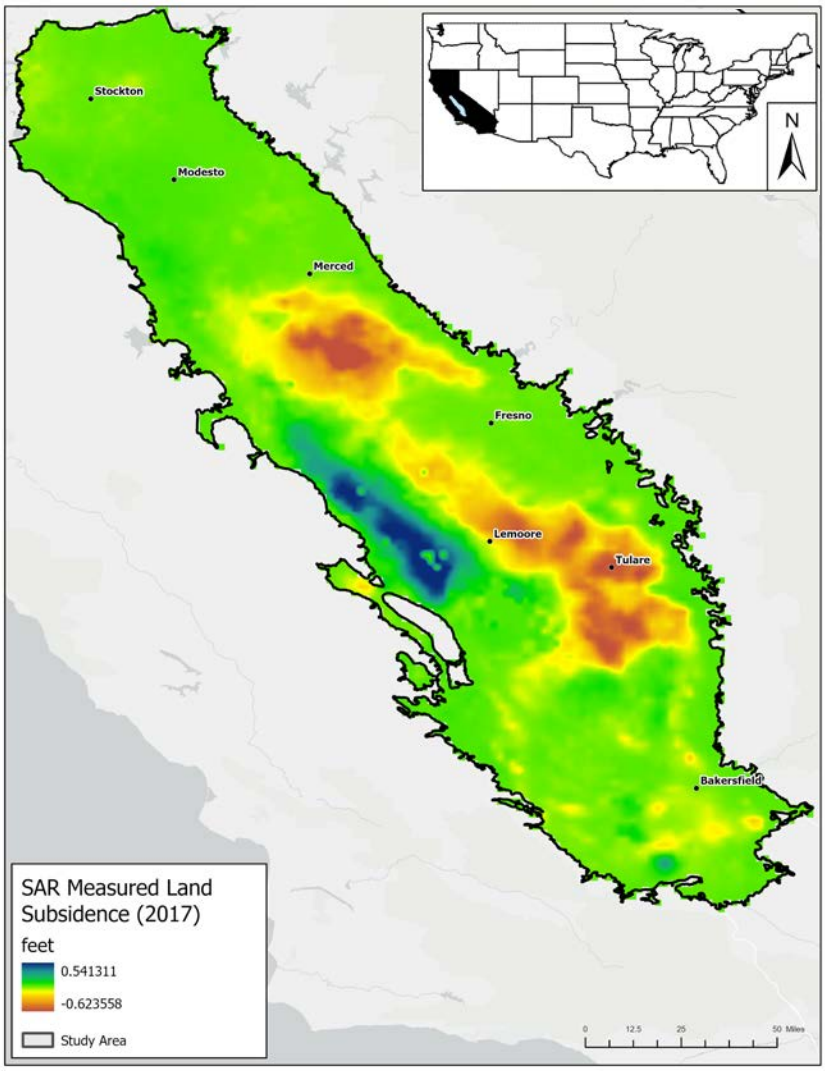


Figure 2. San Joaquin Valley study area map with SAR measured subsidence, 2017

Areas to the southwest of Lemoore are not connected to the hydrologic system of the valley and are part of Kettleman Hills and the Kettleman Dome Oil Field (Harden 2004). While these areas are not part of this study, again as outlined in Figure 2, Kettleman Hills must be mentioned due to the gap or hole that this geologic structure creates in the study area. With this,

all of the light-blue shaded area in Figure 2 shows the AOI of the valley, of which is impacted by subsidence associated with groundwater extraction.

The San Joaquin Valley boundary defines the area of interest (AOI) of this study. This polygon was created in ArcGIS Pro. This feature class was drawn based on the geophysical boundaries of the San Joaquin Valley Kings Subbasin and the Tulare Lake basin watersheds and was guided by groundwater wells the exist with both sub basins. As such, the polygon was drawn with an eastern border of the Sierra Nevada Mountains and the Coast Range mountains on the west. The southern boundary includes the Bakersfield Arch down to and against the Transverse Range. The northern extent is defined by the Stockton Delta. This boundary is ultimately defined by associated watersheds that comprise what it referred to the San Joaquin Valley (DWR 2022).

Figure 2 also shows land subsidence recorded from synthetic aperture radar (SAR) in 2017 within the San Joaquin Valley. Note that negative and warmer colors imply areas of areas of recharge where water may even be seeping to the surface. Positive values and darker blue colors imply areas of greater subsidence or decrease in elevation above sea level.

1.3. Managing Land Subsidence

The amount of subsidence within California's San Joaquin Valley varies over time and is dependent on droughts, infiltration of runoff, and available pore space in the subsurface (Faunt et al. 2016; Jeanne et al. 2019; Poland 1972; Smith and Majumdar 2020). The same may be said anywhere excessive groundwater extraction may be found. Such areas as Shanghai, (Shen and Xu 2011; Xu et al. 2016), Mexico City (Kirwan and Megonigal 2013), Bangkok (Phien-Wej et al. 2006), Iran (Amiraslani and Dragovich 2011; Rahmati et al. 2019), and Las Vegas (Bell et al 2008; Hoffman et al 2001), have undergone large amounts of subsidence. However, the San Joaquin Valley tends to be a textbook example of land subsidence. When assessed at the local

scale, subsidence is based on a multitude of local factors, including overexploitation, water-level drawdown, geology, and water-year type (California Department of Water Resources 2022). With the implementation of California's Sustainable Groundwater Management Act (SGMA) in 2014, a statewide framework was created to assist with the management and use of groundwater that can be maintained without causing an "undesirable result." SGMA describes undesirable results as

> Persistent lowering of groundwater levels, significant and unreasonable reduction in groundwater storage, significant and unreasonable saltwater intrusion, significant and unreasonable degradation of groundwater quality, significant and unreasonable land subsidence, and surface water depletion having significant and unreasonable effects on beneficial uses (Sustainable Groundwater Management Act of 2014 §10733.2).

Groundwater sustainability agencies (GSAs) are local agencies, oftentimes coalitions, for high and medium priority basins that have had significant undesirable results. High and medium priority basins are defined as critically overdrafted basins. With this, GSAs are tasked with developing and implementing groundwater sustainability plans (GSPs) that assist in avoiding undesirable results and help to mitigate overdraft in a twenty-year period (DWR 2022).

1.4. Watersheds and Subbasins

Management of California's groundwater basins was first thought up in the early 20th century. This came as California's population was exploding and rapid growth of the agricultural industry was occurring. Both were not only growing, but also becoming more dependent on groundwater extraction to meet demands (DWR 2020). Within the San Joaquin Valley, groundwater may be found in stratigraphic layers comprised of permeable and porous sediments, the former helps to lead to a high yield capacity (i.e. more easily extracted groundwater). These geologic layers tend to be laterally extensive, but in the early 1950's, California's Department of

Water Resources took the initiative to define geophysical boundaries within existing basins. This soon became known as a subbasin.

California's groundwater systems of interaction are defined by 515 subbasins. The San Joaquin Valley is made up fifteen hydrographic subregions, or drainage basins, which include the San Joaquin Valley and Tulare Basin which in the past decade have been the most impacted by land subsidence due to groundwater extraction (Galloway et al. 1999). These two basins, and the remaining 19 found within the San Joaquin Valley, contain the Tulare Formation, and its three stratigraphic subdivisions which define the primary aquifers.

1.4.1. The San Joaquin River Watershed

The San Joaquin River Watershed is approximately 15,600 square miles and is located in between the Sacramento River Watershed to the north and Tulare Basin Watershed to the south. The San Joaquin River watershed is bordered on the east by the Sierra Nevada Mountains and on the west by the Coast Range mountains. At the heart of this watershed is the San Joaquin Valley - Kings subbasin.

Water flow in the San Joaquin River have been substantially modified by dams and diversions that remove 95% of the water from the river. These diversions cause the San Joaquin River to be dry for more than sixty miles of its course. Some stretches of the San Joaquin receive minimal amounts of agricultural and urban runoff. The Delta Mendota Canal was constructed to replenish water in the San Joaquin River by transporting Sacramento River water to Mendota Pool where it is directed to the San Joaquin River channel and agricultural users.

The land area in the San Joaquin River Watershed is diverse ranging from snow covered peaks to sub-sea level agricultural areas. There are large areas of forest that cover mountain slopes, more than 3000 square miles of agriculture in the valley, and a human population of 2

million people living in the major urban centers of Stockton and Fresno, small towns, and rural communities.

The San Joaquin River is the second longest river in California. It begins in the high Sierra Nevada Mountains and flows approximately 100 miles to the west then turns north flowing for 260 miles where it joins the Sacramento River. Tributary rivers that flow into the San Joaquin River include (from south to north) the Fresno, Chowchilla, Merced, Tuolumne, Stanislaus, Calaveras, Mokelumne, and Cosumnes Rivers (DWR 2022).

1.4.2. The Tulare Lake Watershed

The Tulare Lake Basin is located south of the San Joaquin River watershed bordered on the east by the Sierra Nevada Mountains, on the south by the Tehachapi Range and west by the Coast Range. Major rivers in the Tulare Lake Basin come out of the Sierra Nevada Mountains and include the Kings, Kaweah, Tule, and Kern Rivers. Smaller Sierra Nevada streams include Deer Creek, White River, and Poso Creek as shown in Figure 3.

Figure 3. San Joaquin and Tulare Lake Basin watersheds (NOAA 2022)

Prior to the 19th century the Tulare Lake Basin was characterized by four large lakes. After the basin's tributaries were diverted for agricultural irrigation and municipal water uses, the lakes dried up. These lakes used to cover 800 square miles during wet years. Included in this were large tracts of wetlands that covered close to 625 square miles (Garone 2011). These wetlands had the tendency to periodically spillover into the San Joaquin River watershed. In the modern era, the large lakes and wetlands have been replaced with irrigated agriculture, rural, and large swaths of urban development. Rivers that drain the Tulare Basin do not have a natural surface water pathway out of the watershed. Water moves into and out of Tulare Lake Basin by

precipitation and water diversions through canals. However, the most impacted water source remains groundwater via pumping (DWR 2022).

1.4.3. Subbasin Prioritization and Data

The San Joaquin Valley contains nineteen subbasins for water management purposes, but the valley as a whole is the focus of this study. According to California's Groundwater (Bulletin 118) 2020 updated, the San Joaquin Valley - Kings subbasin uses 2,522,126 acre-ft of groundwater annually. This is ~84% of the water use within this subbasin that is extracted by ~26,684 total wells. That is 17.4 wells per square mile (DWR 2021)! Figure 4 outlines the subbasins that comprise the San Joaquin Valley. Of note is the San Joaquin Valley – Westside Subbasin. This ~970 square mile subbasin is where most of the land subsidence occurs within the valley.

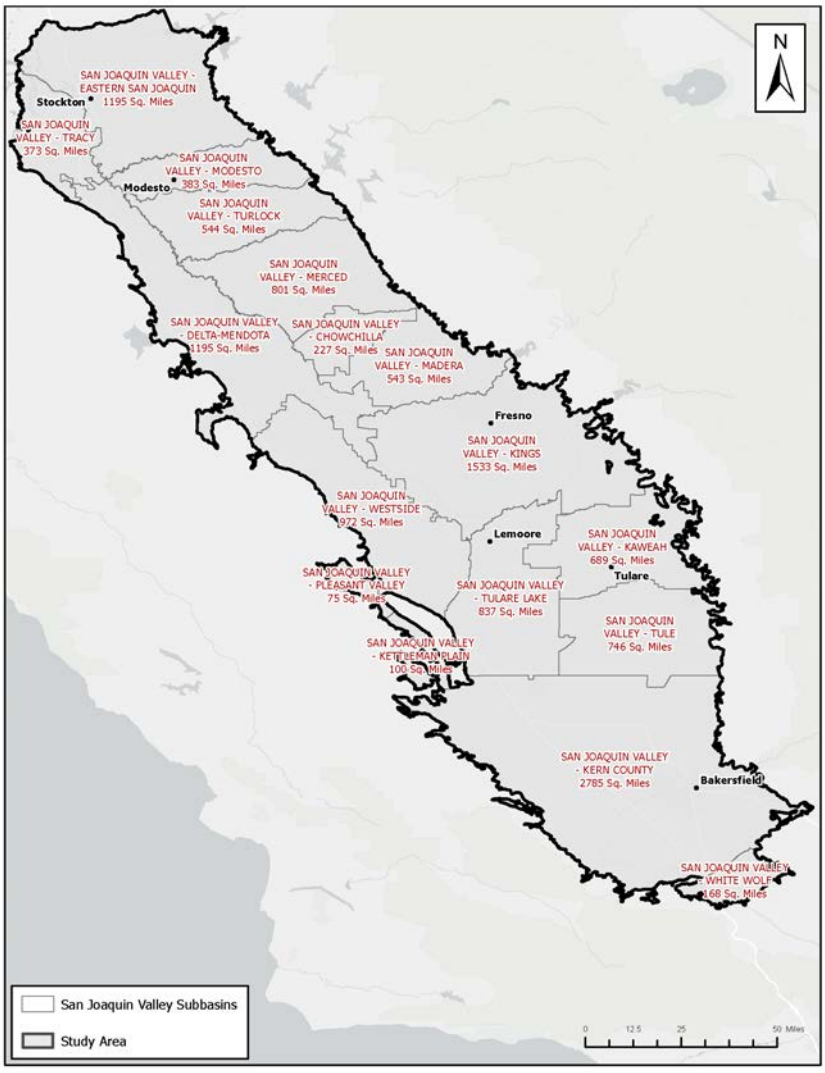


Figure 4. San Joaquin Valley Subbasins

Figure 4 complements Figure 3 and is adapted from the SGMA Basin Prioritization Statewide Summary Table (2022) and outlines similar data for each of the seventeen San Joaquin Valley subbasins, each which have been greatly impacted by land subsidence. It should be noted that according to existing subbasin collations, that are responsible for compiling and managing basin data, and DWR (2018), basin prioritization is based on numerous factors that are identified in CA Water Code §10933. Some factors include current and projected population as well as the

number of existing water wells in the basin. Such rankings then determine if SGMA provisions are applicable within a given basin. The four rankings of very low-, low-, medium-, or high-priority indicate the overall importance of groundwater for each basin and subbasin (DWR 2019). Of the seventeen subbasin within the San Joaquin Valley, thirteen of these subbasins have been marked as high-priority. This tag proliferates from the number of production wells that are present, the number of wells per square mile, and the amount of water extracted from subsurface aquifers. Water extraction is exacerbated by ongoing drought and operates as a catalyst for land subsidence. Table 1 further designates the primary surface stream or river that may exist, albeit more often than not in a dry state, within each subbasin. The area of each subbasin is not something to be missed as that measured surface area reflects the underlying geology which defines the existence of a large subsurface aquifer that is part of the geologic Tulare Formation.

Table 1. SGMA California basin prioritization summary table for the San Joaquin Valley

Subbasin Name	Basin Area (Acres)	Basin Area (Sq. Mile)	Hydrologic Region	Production Wells	Production Wells/Sq. mile	Groundwater % of total supply	Basin Priority
Eastern San Joaquin	764802.7772	1195.004339	San Joaquin River	13144	10.99	0.35	High
Modesto	245252.6544	383.2072726	San Joaquin River	4009	10.46	0.37	High
Turlock	348187.0706	544.042298	San Joaquin River	6606	12.14	0.49	High
Merced	512959.0911	801.4985799	San Joaquin River	5892	7.35	0.66	High
Chowchilla	145574.2985	227.4598414	San Joaquin River	1471	6.46	0.96	High
Madera	347667.3941	543.2303033	San Joaquin River	7059	12.99	0.98	High
Delta-Mendota	764964.8592	1195.257593	San Joaquin River	4041	3.38	0.53	High
Kings	981324.8193	1533.32003	Tulare Lake	26684	17.4	0.84	High
Westside	621823.1784	971.5987163	Tulare Lake	1234	1.27	0.82	High
Pleasant Valley	48195.56034	75.30556305	Tulare Lake	87	1.15	0.94	Medium
Kaweah	441003.9175	689.0686211	Tulare Lake	7385	10.71	0.90	High
Tulare Lake	535869.0664	837.2954164	Tulare Lake	3871	4.62	0.50	High
Tule	477646.4035	746.3225055	Tulare Lake	3360	4.5	0.90	High
Kern County	1782320.811	2784.876267	Tulare Lake	6101	2.19	0.79	High
Tracy	238428.9714	372.5452678	San Joaquin River	2222	5.96	0.03	Medium
Kettleman Plain	63754.60402	99.61656879	Tulare Lake	36	0.36	0.90	Low
White Wolf	107546.2588	168.0410294	Tulare Lake	73	0.43	0.91	Medium

DWR has prioritized groundwater basins based on factors like those previously outlined but emphasis in the San Joaquin Valley has been placed on irrigated acreage and the number of water wells present in the subbasin Water Code §10933 (b)). It may be noted that sustainability managed basins may still be designated as high-priority due to the emphasis further placed on the importance of groundwater within a subbasin and the possibility of degradation of groundwater and undesirable results identified in SGMA.

1.5. Geology of the San Joaquin Valley Groundwater Systems

The San Joaquin Valley is covered by Plio-Pleistocene alluvium sediments from former alpine glaciers (Miller 1989; McPherson and Miller 1990). Sediments eroded from the Sierra Nevada on the eastside and from the Coast Ranges on the westside and were carried via fluvial processes to fill in an asymmetric structural trough. These materials were deposited as alluvial fan, flood-basin, lake, marsh, and deltaic deposits (Miller 1971). Of the 32,000 ft thick sediments, an average of 2,400 ft comprises the aquifer system in the San Joaquin Valley (Page 1961). These sedimentary deposits are comprised of unconsolidated gravel, sand, silt, and clay that define the Tulare Formation which is also the main aquifer with the San Joaquin Valley (Hill 1964). These Tulare Formation sediments hold most of the groundwater reserves throughout the valley as these sediments filled the valley floor from the Temblor range in the west to the Sierra in the east. Additionally, numerous lenses of fine-grained sediments (e.g. silt, sandy silt, sandy clay, and clay) are also present and according to Page (1973) make up over 50% of the total geologic formation and aquifer thickness.

Most of the fine-grained materials have been mapped using geophysical logs, seismic surveys, and drill core throughout the San Joaquin Valley. The most notable lithology is the Corcoran Clay Member of the Tulare Formation that exists along the majority of the westside of

the valley (Bertoldi et al. 1991). The Corcoran Clay Member is a key component of groundwater hydraulics in the valley. Lees et al. (2021) have subdivided the Tulare Formation, the primary aquifer throughout the valley, into three different hydrostratigraphic layers: the unconfined to semi-confined upper Tulare (or upper aquifer), the Corcoran Clay Member, and the lower Tulare (or confined lower aquifer). Figure 5 displays this subdivision. Associating groundwater within the Tulare Formation with the upper or lower aquifer can be difficult. Through the years geologists have used physiography, weathering characteristics, and sediment cores to identify which portion of the formation water wells are drilled into (Williamson 1989).

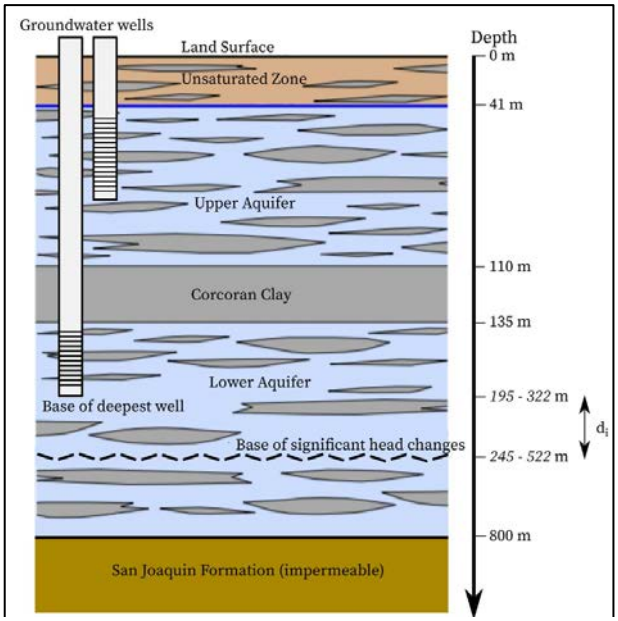


Figure 5. Hydrostratigraphy of the Tulare Formation

Lateral and vertical textural variations among these three subdivisions affect the direction and rate of groundwater-flow as well as the magnitude and distribution of aquifer-system compaction, manifested as land subsidence.

Land subsidence occurs in a particular manner that deals with the rearrangement of fine-grained materials when preconsolidation stress is exceeded (Galloway et al. 1999). This most

often happens when the existing groundwater levels are lower than previous historical lows. Meanwhile, recoverable subsidence is when preconsolidated stresses are not exceeded and hence behave elastically; that is to say, that current groundwater levels remain higher than historical lows (Faunt et al. 2016; Narasimhan and Neuzil 2008; Terzaghi 1923). Such stresses may be caused by land-use changes, ill-managed aquifer recharge, and/or droughts.

The primary concern around groundwater in the San Joaquin Valley is the decrease in overall subsurface storage capacity due to the gradual compaction and sinking of the ground, also known as land subsidence. Subsurface compaction generally tends to occur when large volumes of groundwater are pumped from subsurface pore space faster than natural recharge can replace it. Compaction occurs when sediment is unconsolidated and has high clay content (Davis and Poland 1957; Davis et al. 1964; Poland et al. 1975).

It should be noted that compaction of the Corcoran Clay contributed less than 10% of the total subsidence, and most of the subsidence occurred in the lower aquifer. This is consistent with stress-strain measurements of extensometers that assess the lengthening or stretching of subsurface sediment and geologic rock formations due to downward movement in the subsurface (Poland 1975).

Faunt et al. (2015) discovered that up to 30% of the overall subsidence occurs in the upper aquifer. Faunt et al. (2015) determined that lower aquifer water levels were relatively constant whereas upper aquifer head levels showed an overall decline. This posed an interesting conundrum as it was originally believed that land subsidence was only a problem at locations where the Corcoran Clay is present due to the confined conditions it creates in the underlying lower Tulare Formation (Murray et al. 2018). Figure 6 outlines the extent of the Corcoran Clay

as mapped by Faunt et al. (2015). Notably, the Corcoran Clay thins towards the east side of the valley and designates where aquifer conditions should behave as unconfined.

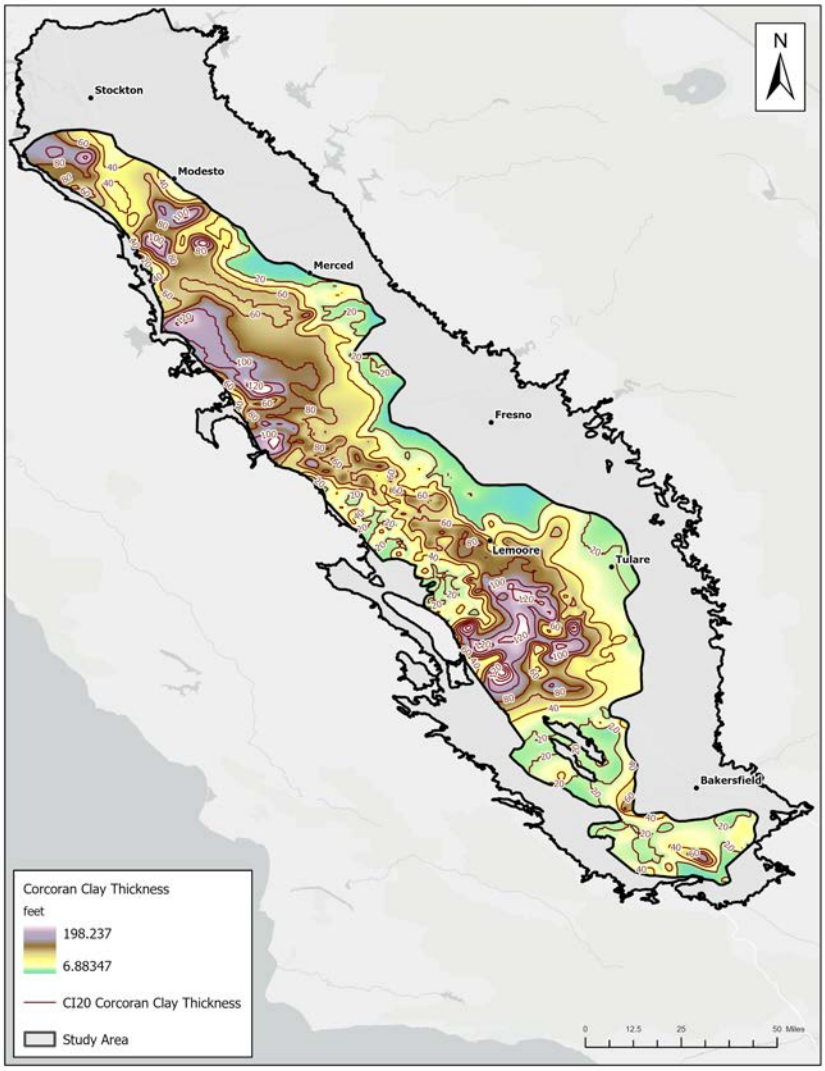


Figure 6. Map of the thickness and extent of the Corcoran Clay

Lees et al. (2021) assessed 65 years' worth of land subsidence to ascertain the depths at which compaction is occurring within the San Joaquin Valley's Tulare Formation. With the work of Lees et al. (2021), it is now understood that substantial subsidence can emanate from within the unconfined-to-semi-confined upper Tulare Formation. This becomes key in understanding

how groundwater and land subsidence modeling might be undertaken within the San Joaquin Valley. Additionally, the findings of Lees et al. (2021) suggest that if heterogeneity is accounted for when spatially modeling the San Joaquin Valley, the presence of the Corcoran Clay—or the lack thereof, may not be as significant as previously thought when it comes to subsidence susceptibility.

The subsurface geology sets the stage for groundwater extraction in the San Joaquin Valley. Unfortunately, excessive groundwater pumping in the San Joaquin Valley has been an ongoing problem since the 1920s. Gradual elevation decline or sinking of the basin's land surface has occurred by as much as 28 feet (8.5 meters). Land-use changes, ill-managed aquifer recharge, and/or droughts compound the problem (Buis and Thomas 2017). The agriculture industry relies on groundwater to support the most productive region in the nation. Hanak et al. (2019) found that the San Joaquin Valley has an annual overdraft of roughly 2 million acre-feet per year. Approximately 30% of the groundwater demand is supplied from pumping groundwater in the San Joaquin Valley in the state of California. This makes the basin the second-most-pumped aquifer in the United States (USGS 2022). Additionally, this makes for ~33 million Californians reliant on groundwater for drinking water or other household purposes. And ~6 million Californians are entirely dependent on groundwater for all water used (DWR 2020).

Pumping at this rate, with no additional wells being added to the system, and maintaining a constant estimate of demand growth, will greatly impact the overall subsurface storage capacity. As groundwater acts as a buffer to drought seasons, DWR (2020) found that groundwater provided 58% of the Californian water supply during the 2012-2016 drought. Groundwater aquifers also play a key role in California's climate change adaptation strategy. If

storage capacity is reduced due to overdraft and continued drought resulting from climate change, then this adaptation strategy also will need to change. However, as many researchers have noted, restoring balance will require a combination of new water supply investments and programs to manage demand—all of which also entail heavy societal and financial costs (USGS 1969; Far and Lui 2015; Faunt et al. 2015; DWR 2022).

As previously mentioned, most water wells in the San Joaquin Valley target the fine-grained sediments of the lower Tulare. With the Corcoran Clay acting as a confining layer, this aquitard tends to deform both elastically and inelastically. Elastic deformation implies a temporary deformation in the sediment and rock of a formation. After the stress, or force, that is causing the deformation is release, the sediment and rock return to their original shape. Inelastic deformation does not reverse and results in permanent change in shape (Twiss and Moores 2007). Land subsidence occurs in a particular manner that deals with the rearrangement of fine-grained materials when preconsolidation stress is exceeded (Galloway et al. 1999). This most often happens when the existing groundwater levels are lower than previous historical lows. Meanwhile, recoverable subsidence (or inflation) is when preconsolidated stresses are not exceeded and hence behave elastically; that is to say, that current groundwater levels remain higher than historical lows. Such inflation cannot occur if the fine-grained sediments have already collapsed, thus making land subsidence permanent. Aquifers cannot have their original storage capacity restored once land subsidence has occurred (Terzaghi 1923; Narasimhan and Neuzil 2008; Faunt et al. 2015).

Understanding the relationship between subsidence and groundwater extraction becomes essential to the future of groundwater and storage capacity in the San Joaquin Valley. The key to understanding falls into several categories of hydrogeologic modeling.

Chapter 2 Related Work

This chapter outlines previous research and work related to modeling groundwater systems and land subsidence. While modeling groundwater systems and extraction methods is a global challenge, the modeling of land subsidence occurs in select areas of the world, with many publications focused on the San Joaquin Valley. Mathematical, analytical, and numerical models have been rendered in 2D and 3D. Recent land subsidence modeling trends have shift from full 3D models to global regression models that then have shifted to local regression models. Examples from other locations on the globe have led the way in assessing the use of geographically weighted regression techniques, including geographic temporally weighted regression models.

Understanding local variation in groundwater levels and subsequent land subsidence has grown more and more important. This study utilizes recently suggested techniques and methods in land subsidence modeling on the basis that not all land subsidence proliferates equally and not all drivers of subsidence are equally distributed. This chapter starts by outlining groundwater modelling techniques and then outlines the differences among models and concludes with an introduction to how geographically weighted regression models have been used to model land subsidence.

2.1. Modeling Groundwater Systems

When it comes to modeling groundwater there are various approaches, methods, and software that are available. Such models are a computational method that are an approximation that representing physical phenomena—in this case, hydrogeologic systems (Barnett et al. 2012). GSAs use many different models to depict groundwater interactions within their respective subbasins. Over the last several decades numerous techniques have been used in attempt to

analyze, predict, and evaluate land subsidence. Numerical and analytical models have been developed over the last 30 years to assist in this area. Many of these models are complex and require many personnel hours to collect and clean data to then generate a model that only groundwater experts may fully understand. Examples of this process include collecting drill core from the subsurface and assessment of recovered core to identify sections for select lab analyses. Lab analyses may include measurement of available pore space, measurement of permeability, and the recording of coarse-to-fine-grain sediment ratios, to name a few. Finally, with the ground truthing of lab analyses from the drill core, alignment of geophysical log measurements of electrical resistivity and conductivity can now be calibrated to properly integrate X, Y, and Z directions of these data into a 3D modelling that can be both expensive and require hours of computer modeling time to be sure results still match the original drill core. Now what happens when a new drill core is taken? Or if a water well is drilled and new geophysical logs exist without drill core to calibrate such measurements to?

A simplified approach of spatial regression can be used to generate predictive models more quickly and accurately while also identifying the most reliable and impactful predictor variables for assessing patterns of land subsidence. In 2009, the USGS produced the CVHM. The aim of this model was to introduce subbasin water managers to how water moves in a hydrogeologic system. This concept would then be complemented by water modelers to make predictions on water moving into and out of the basin. The CVHM has a large list of parameters for managers and modelers alike that must be set before any predictions may be made. The list of data input requirements are related to geology, topography, remote sensing, climate, land use, soils, and chemistry, to name a few. These may come in the form of measured porosity and permeability from a drill core, or the total dissolved solids measured in a water sample taken

from the sub surface. Other data inputs may include annual precipitation, or even predictions of future precipitation wherein a second layer of error may be introduced into the resulting predictions. Due to how quickly each of these inputs can change, while coupled with the amount time it takes to run the model from data input to prediction output, the CVHM has not been updated since 2014. Coincidentally, this is also when SGMA was signed into California law.

In the case of the existing nineteen subbasins within the San Joaquin Valley, conceptual models, mathematical models, and analytical models (and tools) have been utilized. Each of these model types requires different assumptions, datasets, and temporal information. The challenge is that each of these models are used to balance the checkbook of undesirable results. While many are great for assessing and providing rough estimates for a water budget, they are not all the same, and nor should they be. The geology is different. The amount of groundwater extraction (or drawdown) varies throughout each subbasin. Each has a different industrial specialization as well as different population densities that draw on the same groundwater aquifer of the Tulare Formation.

Technologies like the interferometric synthetic aperture radar (InSAR), which is used to measure changes in land surface altitude, have been introduced to reveal, target, and monitor ground deformation from radar images collected by orbiting satellites (Bawden et al. 2003). A radar signal is produced by an orbiting satellite, that radar signal is then reflected back to the satellite to measure the elevation of the area of interest. InSAR images are then created through reference to previously recorded signals at different times and references these against the newly acquired data to create a surface displacement raster (Galloway 2000). Tiltmeter technology has also been used to assess tiny changes in the slope angle or “tilt” of the ground. This is useful in determining the shape or strain of the earth’s crust that results from land subsidence (Fergason et

al. 2015). Global positioning systems (GPS) have also been used to measure geodetic monuments horizontal and vertical changes in land subsidence-prone areas (Sneed and Brandt 2013). As new technologies and data from InSAR, tiltmeter, and global positioning systems (GPS) are implemented to assess land subsidence, the variables of geology, engineering, and groundwater are coming together to complement technology in assessing land subsidence. With this, the simple fact that groundwater drawdown is inextricably linked to land subsidence remains (Sneed 2018). Geology plays a key role as pore space, clay content, sediment grain size, and inelastic vs. elastic deformation contribute to assessing patterns of land subsidence (Galloway and Burbey 2011).

Statistical models, herein coined "quantitative methods," tend to focus on local conditions, but can limit large-scale area assessments (Ali et al. 2020). Quantitative methods dealing with lithology type, surface impermeability, and previously mapped historical subsidence build a much-needed bridge of real-world, boots-on-the-ground application between the field and database. Yet such approaches, much like the CVHM, tend to stagnate and require large datasets if water resource managers and modelers desire newly refreshed predictions (Jeanne et al. 2019). Machine learning, artificial intelligence, and deep learning yield maps that are useful for subsidence prediction and mapping but can still require large training datasets and personnel time.

When it comes to modeling complex geological systems, such as hydrogeologic systems, there tend to be several ways that this may be done. Regardless of the chosen method from among mathematical-analytical, mathematical-numerical, and integrated models, each starts with a conceptual model.

2.1.1. Conceptual Models

A conceptual model is often considered the first step in understanding the groundwater flow system and must occur before a mathematical model can be developed. Conceptual models include a narrative interpretation and graphical representation of a basin based on known characteristics and current management actions. Conceptual models tend to not include quantitative values and are used for conveying complex information in an easy-to-understand way (Castellazzi et al. 2016).

2.1.2. Mathematical Models

Mathematical models, in this context, are designed to simulate groundwater flow or solute transport via solving an equation, or series of equations, that reasonably represent the physical interactions and transport processes in the subsurface. In the case of land subsidence, mathematical models tend to discuss both mechanisms of the effects that cause land subsidence, while also taking a traditional, non-spatial, statistical approach. Mathematical models differ from conceptual models in that they can provide quantitative estimates that can go into a subbasin's water budget. Mathematical models are frequently divided into two categories: analytical and numerical (Bear and Corapcioglu 1981; Yue et al. 2009).

2.1.3. Analytical Models

Analytical models include assumptions that help to simplify complex physical systems. Such simplification may include topographic boundary conditions generally being limited to simple geometric shapes and/or aquifer properties that are often required to be homogeneous and isotropic. The physical configuration of such models is also typically idealized for the purposes of analysis and, therefore, influences related to project geometry are ignored. Often only one component (a measured or simulated value or relationship) of the groundwater system is

evaluated at a time. Such an approach omits the evaluation of potential interactions with other components. In the case of this study, such a method would have the potential to ignore two-way (or more) interactions among pore space, drawdown, and subsidence. Such models are often like balancing a checkbook and include a spreadsheet that utilizes a simple equation to estimate the aquifer drawdown in a single location based on pumping at another location. Such scale-like balancing does not consider the potential influence of heterogeneity in the subsurface or the influence of surface water interactions (e.g. influent streams) (Walton 1979; Ahmed et al. 2020).

2.1.4. Numerical Models

Numerical modeling tools are widely used in groundwater flow and transport analysis to evaluate changes in the groundwater systems caused by changes in conditions due to changes in population and land use, climate change, or other factors. These numerical models allow for a more realistic representation of the physical system, including geologic layering, complex boundary conditions, and stresses due to pumping, recharge, and land use demands. Such a model incorporates complex basin characteristics including significant groundwater withdrawals and/or surface water - groundwater interactions that may be used to estimate when undesirable results may occur (Faunt 2009; Hanson et al. 2010).

Numerical models have come to include Interferometric Synthetic Aperture Radar (InSAR), tiltmeters, and even fiber optic lines buried in the subsurface (Galloway et al. 2016; Ahmed et al. 2019; Guzy et al. 2020). As previously mentioned, these models tend to be complex and while they incorporate predictor variables that represent many possible physical phenomena in the real world, they often include variables that do not bring the most value to the modeling effort. As such, numerical models can easily be biased (Chi and Zhu 2019).

2.1.5. Global Regression Models

Regression is a statistical technique that associates a dependent variable to one or more independent or explanatory variables. A regression model can show whether changes observed in the dependent variable are related with changes in one or more of the explanatory variables (Stapleton 2009). The most common regression technique is ordinary least squares (OLS) by which a linear regression model is established among all data at a global data scale. Variables in an OLS model tend to take on a relationship as outlined in Equation 1.

$$y_i = \beta_0 + \sum k \, \beta_k X_{ik} + \mathcal{E}_i \qquad \text{Equation 1.}$$

Standard regression (OLS) is fixed for the study area. Within Equation 1, *i* represents the observations at a location and *k* represents each designated explanatory variable pertaining to that location. β_0 represents the intercept value; βk is the *k*th coefficient of the independent variable, X_{ik} represents annual groundwater level change and Y_i is subsidence at location *i*.

While OLS is a simple but powerful way to make predictions based on linear relationships, it fails to take into consideration the variance of variables throughout the study area. When areal units like water well locations, street segments, or census blocks are assessed with OLS, assumptions are being made that the observed values at one location are independent of the observed values at other locations. Recall that Tobler's first law of geography outlines the concept that all things are related to each other, but things that are closer in proximity are more closely related than distant things (Tobler 1970). Groundwater extraction does not occur randomly across space as aquifers tend to be confined to a particular area. High-volume water wells will most often be found near other high-volume water wells. This is something that OLS does not take into consideration as such a method looks at all data as a whole, or global scale, rather than as spatially dependent (Harrell 2015).

When a value observed in one location depends on the values observed at neighboring locations, there is a spatial dependence. And spatial data may show spatial dependence in the variables. Why should spatial dependence occur? There are two reasons commonly given. First, data collection of observations associated with spatial units may reflect measurement error. This happens when the boundaries for which information is collected do not accurately reflect the nature of the underlying process generating the sample data. A second reason for spatial dependence is that the spatial dimension of a social or economic characteristic may be an important aspect of the phenomenon (LeSage 2008).

Spatial regression is a common method to predict and quantify spatial patterns (Mitchell and Griffin 2021). Spatial regression methods allow one to model, explore, and investigate spatial relationships that in return can help explain factors that exist behind observed spatial patterns. As such, assessment of independent, or explanatory variables, can lend insight to rates of land subsidence.

2.1.6. Geographically Weighted Regression

Geographically weighted regression (GWR) is a local spatial statistical technique for exploring spatial heterogeneity or non-stationarity. GWR is powerful when relationships between X and Y vary by locality and thus is commonly used to predict spatial variation of a relationship with the dependent variable—in this case, land subsidence (Matthews and Yang 2012).

GWR constructs a separate OLS equation for every location within the project AOI. It incorporates both dependent and independent variables that fall within a searching bandwidth of each location. This is often termed as local regression as not all variables and their location are being assessed at the exact same time.

GWR weighs observations in association with their proximity to what is termed *i*. Following suit with Tobler's First Law of Geography, observations that are found closer to *i* have a stronger influence on the estimation of the parameters for that location. Equation 2 outlines how the relationship between the dependent variable (land subsidence) and covariates (e.g. groundwater level) are established through homogeneity. Homogeneity in this scenario is defined as sharing the same relationship between the dependent variable and some, but not all, covariates. The ability for GWR to take into consideration different neighborhoods and how they change is what makes GWR fit for local regression. Here *r* represents each regime or regional regression will be performed in. For example, r = 1, 2…m (Thapa and Estoque 2012).

$$Y_r = \sum \beta_r X_R + \mathcal{E}_r \qquad \text{Equation 2.}$$

This results in geographically weighted regression (GWR), as represented in Equation 3 and by which a new regression model is generated and varied at each point *i*. β0 (u_i, v_i) designates the coordinates of the i-th point spatially. βk (u_i, v_i) represents a realization of the continuous function at point *i* in space (Thapa and Estoque 2012).

$$Y_i = \beta_0(u_i, v_i) + \sum k \beta_k (u_i, v_i) X_{ik} + \mathcal{E}_i \qquad \text{Equation 3.}$$

Geographically and temporally weighted regression (GTWR) considers the temporal aspect of each independent variable as it influences the dependent variable through space and time (Ali et al. 2020; Chu et al. 2021). Geographical weighted temporal regression is an

extension of GWR by which the additional dimension of time (t) is added to the already spatially aware regression model (Miller and Shirzaei 2015). Equation 4 shows the formulation of GTWR.

$$Y_i = \beta_0(u_i, v_i, t_i) + \sum k \beta_k (u_i, v_i, t_i) X_{ik} + \mathcal{E}_i \quad \text{Equation 4.}$$

In this case, $\beta_0(u_i, v_i, t_i)$ defines the intercept that now incorporates both location and time at i. Meanwhile, $\beta_k(u_i, v_i, t_i)$ represents the estimated coefficient at each spatio-temporal observation (i) as it relates to variable k. As with the GWR formula, X_{ik}, for example, will represent annual groundwater drawdown (or other statistically significant independent variables identified in the ESDA process) and Y_i represent the annual groundwater drawdown and subsidence, respectively. In this study, k represents additional explanatory variables such as well depth, storativity, area, well vintage, well completion length, confining clay layer thickness, and lithology.

GTWR is used to help control spatial errors and is useful in identifying areas within the basin where preconsolidation stresses are exceeded—something that geologic engineers and civil engineers alike are concerned about. The relationship between groundwater drawdown and subsidence via GTWR has become an accepted method for assessing elastic vs. non-elastic deformation (Faunt et al. 2016; Guzy and Malinowska 2020).

Ali et al. (2020) set time-variable $Y_t = (Y_{1,t}..., Y_{n,t})^T$ and X_t into a corresponding matrix of covariates. They go further to set an estimated coefficient matrix, $\beta_t(u_i, v_i)$, at every time period. Equation 5 shows the derivation of this effort.

$$\widehat{\beta_t}(u_i, v_i) = [X_t^T W(u_i, v_i) X_t]^{-1} X_t^T W(u_i, v_i) Y_t \quad \text{Equation 5.}$$

With spatial autocorrelation a primary concern in spatial regression models, control for a spatial error model is likely to be needed (Chi and Zhu 2019). GWR enables researchers to measure and visualize variations in relationships that are unobservable in global, aspatial model, while also minimizing biases due to spatial errors (Fotheringham et al. 2002).

Due to challenges associated with non-uniform hydrogeologic layers, spatial variance in land subsidence made the subsidence-drawdown function of Ali et al. (2020) and Chu et al (2018) perform poorly when assessed with OLS techniques. This is largely attributed to the subsidence-drawdown method being impacted by heterogeneity (Jiang et al. 2019; Sundell et al. 2019). To overcome this, Ali et al. (2020) chose to utilize geographically weighted regression (GWR) to implement a linear regression model comprised of spatially varying relationships (Fotheringham et al. 2003).

The benefits of GWR have recently been identified by researchers involved with land subsidence forecasting due to the ability to assess variable coefficients and spatial nuances among variables throughout the area of interest (Huang et al. 2010; Ali et al. 2020, Chu et al. 2021). While authors have proved GWR to be a powerful tool of prediction and mapping, it has yet to be utilized in areas of severe groundwater depletion such as Shanghai (Sen an Xu 2011; Xu et al. 2016), Mexico City (Kirwan and Megonigal 2013), Bangkok (Phien-Wej et al. 2006), Iran (Amiraslani and Dragovich 2011), Las Vegas (Bell at al 2008; Hoffmann et al. 2001), and the San Joaquin Valley, CA (Faunt et al. 2016; Jeanne et al. 2019). However, and as previously mentioned, GWR has been successfully used in several areas of excessive groundwater exploitation, including the Choshui River alluvial fan (Chu et al. 2021) and in Changhua and Yunlin counites, Taiwan (Ali et al. 2020).

As GWR does not assume relationships among variables vary across space, but rather identifies if there is variability across space, the technique makes it a unique and appropriate way to estimate geological and hydrogeological phenomena due to the non-isotropic, physical nature of the subsurface. The challenge with GWR is that it is not a tool for defining variables of influence and thus should be complemented with exploratory spatial data analysis (ESDA) techniques. ESDA enables one to correlate specific variables to a location while considering the values of the same variable within the neighborhood. Although already mentioned, this approach does in fact define spatial autocorrelation as ESDA goes about describing the presence (or absence) of spatial variation among variables (Haining et al. 1998). ESDA will allow one to identify the most influential variables that may then be used in the GWR process to regression on (Matthews and Yang 2012).

Ali et al. (2020) utilized groundwater level observations throughout 2015 to track and calculate monthly groundwater level change. Building on the geoengineering concept that fine-grained sediments tend to inelastically deform and result in land subsidence (Galloway and Burbey 2011), Chu et al. (2021) established reasonable estimates of land subsidence linked to groundwater level change from OLS regression models. This went to establish a statistical approach to assessing groundwater draw down and its relation to land subsidence. It should be noted that numerous authors had already hinted at this through observation and assessment among geotechnical consulting companies in practice, but few had published on this mathematical relationship (Chu et al. 2018; Narasimhan and Neuzil 2008; Sneed 2018; Terzaghi 1923).

Chu et al. (2021) built on the use of local regression technique via the use of groundwater level changes to predict spatial patterns of land subsidence. The work of Ali et al. (2020) and

Chu et al. (2021) included five hydrostratigraphic layers that are comprised of gravel, sand, and clay. This approach of including sediment grain size follows much of what Faunt et al (2015) has done with the CVHM. Aside from geological and hydrological variables, Ali et al. (2020) and Chu et al. (2021) also utilized land subsidence measurements from hundreds of leveling stations to generate land subsidence maps.

Ali et al (2020) took the approach of regressing a two-variable linear model: measured land subsidence extracted from InSAR as the dependent variable and annual change in groundwater levels from thirty-six monitoring wells as the independent variable. They then went on to assess the performance of a linear model. After this, GWR was introduced to better understand the distribution of model coefficients. These two global and local models were compared yielding R^2 values of 0.34 and 0.93 respectively. Ali et al. (2020) then took the aggregation of their variables from 2007 to 2017 temporally regress the variables. The R^2 value yielded with a geographically temporally weighted regression model was 0.94, demonstrating very little improvement from the GWR models.

Chu et al. (2021) followed the same exploratory techniques as Ali et al. (2020). The difference between these two studies was that Chu et al. (2021) chose to model each of the four aquifer layers present in the Choshui River Basin, China. Their methods did not change, as they developed OLS, GWR, and GTWR models for based on the presence of aquitards separating these four primary aquifers. The change made in aggregation to a finer, aquifer scale by Chu et al. (2021) showed a slight improvement on the R^2 values, with an average of 0.28 for OLS, 0.97 for both GWR and GTWR. In the end, both sets of authors concluded that without requiring a large and detailed hydrogeologic model, or measurements needing extensive calibration (e.g.

geophysical logs calibrated to drill core), spatial regression models utilizing only two variables offer reasonable insight into patterns of land subsidence.

The use variables in the form of sediment grain size, sediment type, and groundwater draw down established the expected relationship between predictor variables and dependent variable (land subsidence), but also allowed a working knowledge of the geology and engineering principles to help guide groundwater spatial regression modeling. This approach emulates Ali et al. (2020) and complements Chu et al. (2018). Chu et al. (2021) found that the subsidence-drawdown relationship is nonlinear.

Since 2015 many authors have utilized Geographically weighted regression (GWR) and spatiotemporal techniques as a means of achieving a faster, yet highly accurate predictive model. This has led to a means of modeling hydrogeologic systems as they pertain to land subsidence (Ali et al. 2020; Hung et al. 2016; Fotheringham et al. 2015; Chu et al. 2021). Such techniques have yielded strong results with a root mean square error (RMSE) much lower than OLS methods as well as a much higher coefficient of determination. These techniques can identify spatial variability among explanatory variables while honoring variables' ability to represent geological and engineering phenomena (Burbey 2005).

2.2. Modeling in the San Joaquin Valley

Modeling groundwater and land subsidence in the San Joaquin Valley has predominantly occur from researchers at the USGS. As mentioned throughout this chapter, the CVHM of Faunt et al. (2009) has been the primary instrument of prediction. The CVHM is comprised of four different models: the geospatial database, the texture model, the MODFLOW simulation model, and the numerical model.

The CVHM utilized a geospatial database to compile, manage, and store the large volumes of data for the CVHM. Much of the analysis was conducted in a GIS to generate and visualize outputs. However, there is another the texture model is the geologic 3D grid cell model that combines hydraulic properties of the subsurface with borehole drill logs, drill core, and geophysical tools. In truth, this has been a fully functional geologic properties model. Something that takes a great deal of time and effort to put together and validate.

The next portion of the CVHM comes from MODFLOW, which is a hydrologic software that the USGS developed to model the flow of water over the landscape. This modelling effort accounts for volumes of water coming into the system (i.e. precipitation) and water leaving the hydrogeologic system (e.g. groundwater extraction). Much like a geologic model, MODFLOW inputs can take a large amount of time to collect and validate. As an example, the most recent model is based on water years 1962 to 2003, showing the difficult researchers and modelers have had in maintaining this portion of the CVHM (Faunt et al. 2015).

The numerical model portion of the CVHM is designed to incorporate all the outputs for the previous models to output predictions, including predictions of land subsidence. However, it goes without saying, that if one portion of the model is dated or other models' performance are not to par, all of the results may be further biased.

There have been engineering consultancies that have studied land subsidence in regions throughout the valley. The catch is that knowledge of their work comes by word of mouth and is not readily available to the public. It is for this reason that the CVHM has been the default model that researchers and water managers alike have turned to. However, since 2020 Stanford University has taken an interest in modeling land subsidence in the valley. In a publication from Lees et al. (2021) utilized subsidence rates and hydraulic head, or the measurement of pressure

above a set vertical datum, to create and validate a one-dimensional model. Their methods consisted of the use of subsurface groundwater flow equations for permeability and clay compaction equations for a small area in the Tulare Lake Subbasin. These methods are not new as the equations were developed by Helm (1975). In the end, Lees et al. (2021) were successful in simulating up to twenty-five feet of land subsidence; however, this approach is limited to small, survey sections.

Current researchers' models take a large amount of time to collect, synthesis, and calibrate data inputs. At the same time, both methods have spatial or other dependencies that restrict a quick turnaround for land subsidence predictions. The methods explored in this study assess spatial regression models that are grounded in geologic and engineering principles while providing methods that are easy-to-update without the loss of accuracy.

Chapter 3 Methods

This chapter provides an overview of the research design, including methodology, data descriptions, and data preparation steps. The last half of this chapter discusses the steps taken to implement spatial regression techniques in the form of generalized linear regression and geographically weighed regression. This chapter ends with an overview on how model performance is assessed for each spatial regression model and the accompanying results of coefficients, R^2 values, AIC values and interpolated rasters from predicted values sourced from each spatial regression model.

3.1. Research Design

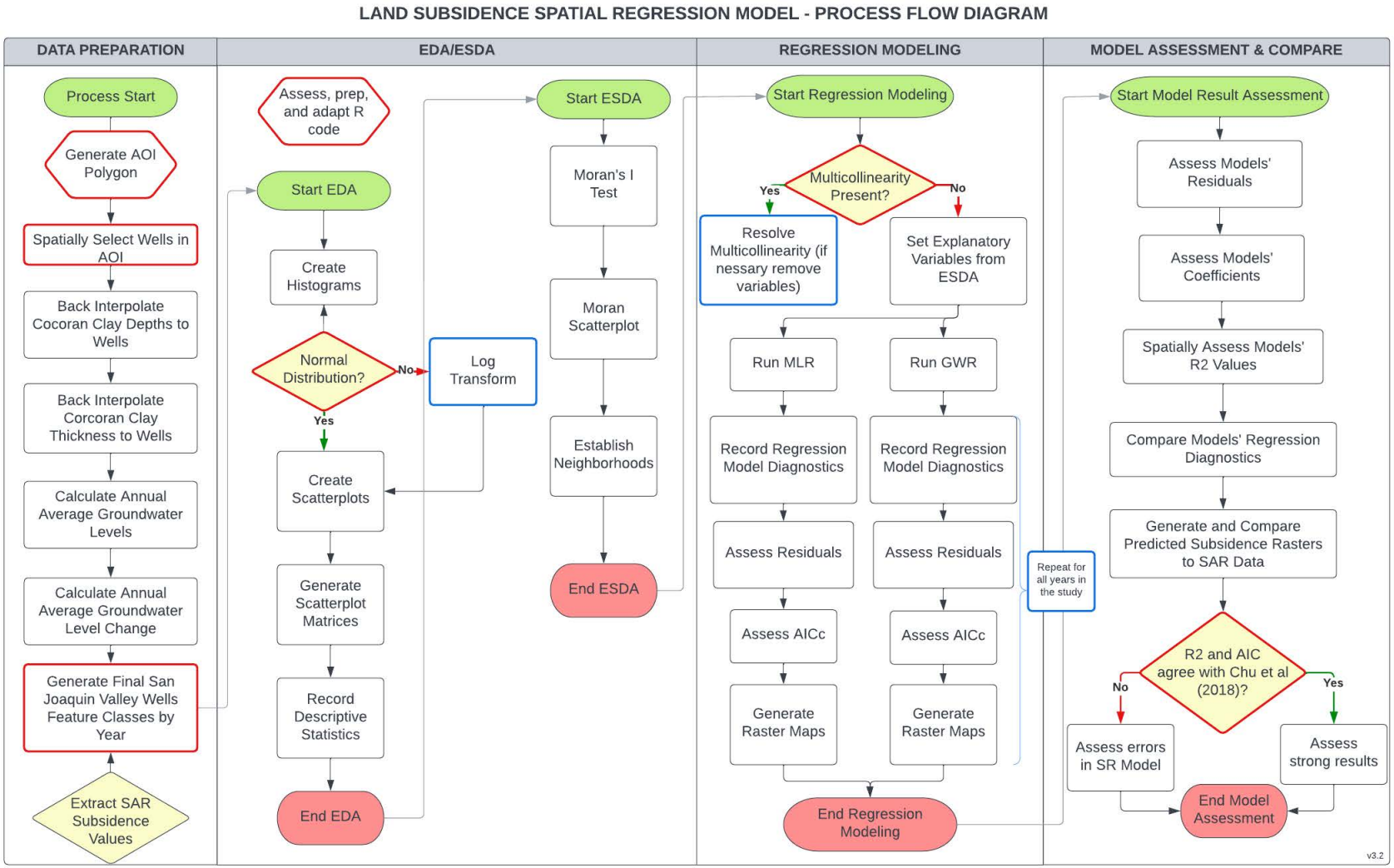
Following several key techniques from Ali et al. (2020), this study builds a spatial regression model to assess spatial patterns of land subsidence. To create a simple, but accurate regression model, ESDA is used to identify which independent variables best assist in predicting patterns of land subsidence in the San Joaquin Valley, CA. Ali et al. (2020) and Chu et al. (2021) utilized geographical weighted regression (GWR) to predict spatial patterns of land subsidence, while accounting for local effects in space and time, (Fotheringham et al. 2015; Huang et al. 2010). As this study goal was to establish a more simplified but accurate model than what currently exists with the CVHM, the temporal aspect is focused on 2015 to 2021. This chosen time frame is where well records and information are the most complete and do not have the errors that pre-2000 data tend to exhibit.

An annual aggregation was chosen as annual synthetic aperture radar (SAR) data are available from DWR within the San Joaquin Valley. These data start in 2015 and are current today. Having this dataset in place set the stage for the creation and testing of a productive spatial regression model. The final step was in this process was to compared model outputs with

the existing SAR data for a qualitative measurement of model accuracy. Additionally, the spatial regression rasters generated are subtracted from the exist SAR rasters to generate a delta map of differences between actual and predicted land subsidence.

Figure 7 provides step-by-step details on the methodology and workflow followed in this study. This process flow diagram (PFD) gives a step-by-step look at prerequisites that were met before GWR was implemented. Some key takeaways come in both shape and color on the PFD. Green filled shapes designate the start of a process, while red filled shapes designate the end of a process. Any shapes that are outlined with a bold red line indicate that the current step must be completed before moving on to the next step in the process; in short, a decision had to be made.

Figure 7. Spatial regression process flow diagram (PFD)



Yellow filled diamonds, that are also outlined in red, are decision points. This means that a choice must be made to generate a product. One such example is the extraction of SAR values from existing rasters. To further elaborate on this example, should the extracted SAR raster values be averaged over a certain number of cells, or should the first encountered value be used? Such decisions were made based on previous works with the extraction of raster values being conducted at the well level.

Blue outlined shapes indicate where choices were made before further analysis may be done. One may note that the distribution of each variable is conducted and assessed for a normal, Gaussian distribution. From here all non-normal data distributions would undergo a log transformation. Note that there are both a decision node (outlined in red) and an action or function node (outlined in blue) noting choices and actions that needed to be taken in this study. An assessment of the data must be made and a choice to proceed or make a log transformation of the data must be made.

The study's AOI was determined based on the well data, basin prioritization and surrounding geophysical mountain ranges that define the San Joaquin Valley. From here, wells were spatially queried and identified based on be their existence within the AOI polygon. Raster values pertinent to depth to the Corcoran Clay, Corcoran Clay thickness, and even the subsidence values for each year were extracted to the well level. Issues stemming from the modifiable areal unit problem (MAUP) must be acknowledged. Due to the way that groundwater extraction wells and their accompanying data are aggregated in this research, there is no objectively recognizable way to reclassify these data without results being impacted. These impacts are often referred to as the "zoning effect" and the "size effect" (Bolstad 2016). To mitigate the negative impacts of MAUP, authors dealing with well data have looked to establishing neighborhoods to associated

well-to-well interactions through Moran's I for Spatial and Autocorrelation (Chu et al. 2021; Ali et al. 2020).

As already emphasized for this study, the choice was made to continue a well-by-well association of data. This means taking the individual well back to its basic unit of measure and minimizing large impacts from MAUP. Well level assessment helps to avoid issues associated with aggregation and should coincide with local variation trends leading to the preferred method of areal units for assessment. While this might hint at Ecological Fallacy, the approach taken in this study follows previous publications and avoids dependence on any single set of aggregate-level mapping units through incorporation of information from many different datasets (Tuson et al. 2020).

From the existing well records, the annual average groundwater levels were calculated and incorporated as a new data field alongside multi point extraction values. Exploratory data analysis was conducted to better understand the state of the data, and as previously mentioned, if data demonstrated a non-normal distribution, these were to be log transformed. Traditional statistical analyses were conducted in the form of descriptive statistics, as well as scatterplots and histograms.

As Figure 7 further outlines, exploratory spatial data analysis was conducted. And as previously mentioned, running a Moran's I test was key to help identify optimal search distances and variance in neighborhood ranges. From here stepwise regression was conducted to identify key independent variables as well as to preliminarily assess how the data perform within an OLS model. Residual plots were generated, and linear fits were attempted.

Emphasis was then placed on variables that did not demonstrate multicollinearity and those that were determined to contribute to the regression model. From here the identified

independent variables were entered into the generalized linear regression (GLR) spatial statistics tool in ArcGIS Pro. Coefficients from this process were assessed alongside R^2 and AIC values. Utilizing the predicted land subsidence values from MLR, IDW interpolation methods were used to generate a quick raster of predicted values. This will be discussed later, but it is acknowledged that this step may have introduced a second layer of error, but this step does follow the works of Chu et al. (2021) and Ali et al. (2020). The same steps were repeated for geographically weighted regression. Similarly, and for comparison, R^2 and AIC values were recorded.

All model performance values were checked against Chu et al. (2018)'s results from subsidence modeling using SAR data. However, emphasis on spatial regression models not having been tested in the San Joaquin Valley is taken into consideration especially as there were more independent variables present in this study that came from larger datasets.

3.2. Data Preparation

This section will cover the selection of data for use with a spatial regression modeling, data formatting for processing in a spatial regression model, and execution of spatial regression in the form of Geographically weighted regression.

3.2.1. Data Description

Table 2 outlines the data used in this study and each is respectively described later in this section. Each dataset is available online via the USGS, CA Department of Water Resources, CA Sustainable Groundwater Management Act portal, and/or the California Open Data portal at no cost. Additionally, several datasets that are key to geology and hydrogeology studies can take on several forms. As an example, one dataset is defined by whether perforations are greater than or less than the top and base depths of the Corcoran Clay member of the Tulare Formation. As noted in the geologic background of the Tulare Formation, knowing if a well is pumping

groundwater above or below the Corcoran Clay is essential. This again defines the difference between a confined vs. an unconfined aquifer. However, the two key datasets are the SAR Annual Vertical Displacement raster data and the Annual Groundwater Well Measurement Stations feature class (here after referred to as the SJV Wells). These two datasets ultimately define the spatial and temporal points of control related to this study.

Table 2. Summary of datasets

Dataset Name	Description	Format	Source
SAR Vertical Displacement Annual Mosaic	Vertical displacement in high-use groundwater basins. Land subsidence values for the selected timeframe as measured by SAR	Raster	DWR
Annual Groundwater Well Measurement Stations (SJV Wells)	Groundwater level time series measurements in CA. Well locations; groundwater table level assessment; total depth of water wells	Vector (point)	DWR
Corcoran Clay Base Depth	Stratigraphic base of the Corcoran Clay Member of the Tulare Formation. Separate Upper and Lower Tulare water wells	Vector (polyline)	USGS
Corcoran Clay Extent	Spatial extent of the Corcoran Clay as mapped by the USGS from subsurface drill core and outcrops	Vector (polygon)	USGS
Corcoran Clay Thickness	Thickness of Corcoran Clay in feet. Separate Upper and Lower Tulare water wells	Vector (polyline)	USGS
Well Completion Reports	List and details of all wells of record with completion depths and dates. Well locations; well completion intervals	Vector (point)	DWR
Percent Coarse-to-Fine-Grained Material	Well log descriptions of % coarse-grained material and % fine-grained material. Fine-grained and coarse-grained material identification per well	ASCII	USGS

3.2.2. Data Sources

Data engineering (i.e. data preparation) was a large component of this project. While this is not unfamiliar for those who publish and work with spatial data, it should be noted that several datasets in this study will be a first of their kind for the San Joaquin Valley. These data are derived from existing datasets available from the United States Geological Survey and the California Department of Water Resources. Even though each stem from publicly available data, that are specific to the San Joaquin Valley, CA, these data have not been brought together in a way that they may be used to predict land subsidence. Of special note would be those water production wells that are completed in the upper Tulare (unconfined aquifer) and those completed Lower Tulare (confined aquifer). How these two specific datasets were created is outlined in their respective subsection pertinent to dataset creation.

As an additional example, some datasets are temporally aggregated. One such dataset is the groundwater wells themselves. This dataset contains groundwater level measurements through time. It contains seasonal and long-term groundwater level measurements collected by the CA Department of Water Resources, as well as GSAs. These data come from measurements that are acquired twice a year. This twice-per-year measurement is meant to capture both high and low values of groundwater elevations as well as seasonality in the San Joaquin Valley's groundwater system. These data were used to generate an average annual groundwater level. Ideally, these data would be used to generate a mean monthly water level dataset, but due to the frequency of surveys throughout the valley, an annual mean is calculated (Ali et al. 2020).

From here the mean annual data were used to calculate the average annual groundwater level change. The relationship between the independent variables and the dependent variable of annual land subsidence are then used to establish a spatial regression model (Ali et al. 2020; Chu et al. 2021).

Specifics around each dataset and associated variables that proliferate from them are outlined in the next sections. These datasets come in the form of spatially queried well locations, and even a designation of vertical separation between confined and unconfined portions of the Tulare aquifer.

3.2.3. San Joaquin Valley Wells

These data were the core of this study and are derived from the SJV Wells dataset courtesy of the California Department of Water Resources (DWR). These data came from DWR as a shapefile that included attribute table fields consisting of well names, well total depths, in feet below ground surface (bgs), and a history of water levels (bgs) from the early 1900's to today.

This dataset was spatial queried based on wells located in the San Joaquin Valley Boundary (AOI). Wells were spatially queried based on the AOI and were exported as a new feature class, thus removing all wells outside of the AOI. Furthermore, depth values that are null or zero were filter based on a Definition Query as such records provide no value to this study and represent a permitted but incomplete (i.e. not drilled) water production well.

3.2.4. Corcoran Clay Base Depth

These data are a contour set of subsurface values for the base of the Corcoran Clay. This dataset came from the USGS and is a culmination of work from the water industry, the USGS, and private industry. Each of the depth measurements have been identified in drill core, geophysical logs, or in seismic surveys and have been used to map the extent and the depth of the bottom of the confining clay layer that define the confined and unconfined portions of the Tulare aquifer.

These contour values were converted to a raster via the Topo to Raster (Spatial Analyst) tool in ArcGIS Pro. The output cell size was set to 100ft. All other optional tool parameters were set to their default. From the generated raster the Extract Multi Values to Points (Spatial Analyst) tool was used to extract all raster data, in this case the Corcoran Clay Base Depth, into newly created data fields within the San Joaquin Valley Wells feature class.

3.2.5. Corcoran Clay Thickness

These data are from a raster that represents the thickness of the Corcoran Clay. This dataset came from the USGS. Much like the Corcoran Clay Base Depth data, this raster for clay thickness is comprised from work conducted in the water industry, the USGS, and private industry. Each of the thickness measurements have been identified in drill core, geophysical logs, or in seismic surveys and have been used to map the extent and the thickness that is formed between the top of the clay and the base of the clay define the confined and unconfined portions of the Tulare aquifer.

Also like the Corcoran Clay Base Depth data, the Extract Multi Values to Points (Spatial Analyst) tool was used to extract all raster data into a newly created data field within the San Joaquin Valley Wells feature class.

3.2.6. Wells Completed in the Upper Tulare; Wells Completed in the Lower Tulare

This dataset was created from a combination of spatial queries and assessment from the Corcoran Clay Base Depth and the San Joaquin Valley Wells datasets. While other researchers have implied the presence of aquitards, such as clays, can have a large impact on land subsidence, no San Joaquin Valley publications have attempted to separate out the upper Tulare and lower Tulare components of the aquifer system. This study intends to bridge this gap between academic publications and the actual practice of geology in the San Joaquin Valley.

By assessing if the top and base perforations of a well are < the Corcoran Clay base depth (bgs), wells were assigned to the upper Tulare. This implies that all groundwater exploitation and extraction was and is occurring from within the upper Tulare Formation for these wells and hence are part of an unconfined aquifer. A binary approach was taken to assign wells to the upper Tulare ("0"). These values were then assigned to a newly created data field within the SJV Wells Attribute Table in ArcGIS.

The same approach was taken and if the top and base perforations are > the Corcoran Clay base depth (bgs); thus, the wells were assigned the lower Tulare. This implies that all groundwater exploitation and extraction was and is occurring from within the lower Tulare Formation, or a confined aquifer, for these wells. A binary approach was taken to assign wells to the lower Tulare ("1"). These values were then assigned to a newly created data field within the SJV Wells Attribute Table in ArcGIS.

If the top and base perforations were found within both the upper and lower Tulare Formation, or no Corcoran Clay depth value exists, then the aquifer system is assumed to be unconfined where groundwater extraction is commingled between both the upper and lower Tulare aquifers. This also implies that there is no seal between the upper and lower aquifers and hence no stratigraphic differentiation is needed for such wells.

It should be noted that each of the data in this study were created from a serious of spatial queries, spatial joins, average calculations, and point extraction exercises that resulted in a combination of data to be studied in a unique way. Each of the values combined and assessed do not exist in a singular USGS database nor have they previously been combined into a single dataset by DWR. The reason for neither agency having such a dataset could be that the depth to the Corcoran Clay has been maintained by one agency and the location of all water production

wells in the San Joaquin Valley have been maintained by a separate agency. The creation of this dataset is one aspect that is different from other models that do not incorporate distinct upper and lower stratigraphic variation within the Tulare Formation. However, Ali et al. (2020) did in fact separate their study into five distinct hydrostratigraphic layers, while Chu et al. (2021) assessed five distinct hydrostratigraphic layers. In both publications, some of these hydrostratigraphic layers are impacted by the presence of aquitards (i.e. thick clay layers).

3.2.7. Annual Average Groundwater Level

These values of groundwater measurements were added to a newly created field in the Wells in the San Joaquin Valley Attribute Table. These values were derived from the recorded measurements provided in the original source data (SJV Wells) and were calculated by summing the survey values for each year and dividing them by the number of surveys each well had in an annual timeframe. It is acknowledged that this approach is a level of aggregation that may impact results, but once again follows the work of Ali et al. (2020) and their aggregation as monthly average groundwater level.

3.2.8. Annual Average Groundwater Level Change

These values are part of a data field that was created in the Wells in the San Joaquin Valley Attribute Table. This field was created by taking the values from the annual average calculation and subtracting each average from one year to the next (e.g. 2016 average value – 2015 average value). This data field may be positive, implying inflation or recharge of the aquifer or it may be negative, implying loss of groundwater and subsequent lowering of the water table. It is again acknowledged that these data contain a level of aggregation that may impact results, but once again follows the work of Ali et al. (2020) and their aggregation as monthly average groundwater level changes.

This data field and the Annual Average Groundwater Level are the two key data fields for assessing spatio-temporal patterns of land subsidence (Chu et al. 2021).

3.2.9. Well Completion Percent Fine-grained vs. Coarse-grained Sediment

As with the previous datasets, this dataset was turned into another Attribute Table field within the Wells in the San Joaquin Valley feature class. This field defines the percentage of fine-grained to coarse-grained material that has been recorded in completed and logged water production wells. Table 3 outlines the Modified Wentworth Scale that geologists use to define grain size of clastic rocks and sediment. This system allows one to classify sediment according to the size of particles (Blatt, Middleton, and Murray 1972).

Table 3. Modified Wentworth Grain Size Scale

Diameter (mm)	Particle	Sediment	Rock
<1/256	clay	mud	claystone, mudstone, shale
1/256 to 1/16	silt		siltstone
1/16 to 2	sand	sand	sandstone
2 to 4	gravel	gravel	
4 to 64	pebble		conglomerate (rounded)
64 o 256	cobble		breccia (angular)
>256	boulder		

This dataset is defined by the well-log-texture ASCII that comes from the USGS. This dataset contains associated depths and percentages for a lithologic model created from drill core throughout the San Joaquin Valley as well as geophysical resistivity logging upon the wells being drilled (USGS 2016). These data are added to the Wells in the San Joaquin Valley feature

class by field calculations where values found with the range between the top perforation and the base perforation are averaged. This yields two new Attribute Table fields. One that contains the percentage of fine-grained sediments within the completed zone, as well as the average percentage of coarse-grained sediments within the completed zone.

It should be noted that this dataset if complex and has a large 3D component. While the goal of this project is to create a quick but accurate way identify spatial patterns of land subsidence, it is not meant to be fully function 3D geological or lithological model. This approach does further introduce challenges associated with MAUP; however, previous researchers have been able to take depth averages to generate accurate geological models that are fit-for-purpose (Galloway et al. 2011; Faunt et al. 2015; Ali et al. 2020; Chu et al. 2021). The alternative is to build a fully function 3D model based on variogram methods which is well outside of the scope of this project.

3.2.10. Well Completion Length

This dataset is further derived from the source DWR data and is tied to the SJV Wells. This field that was added to the feature class Attribute Table of the SJV Wells is defined by taking the top perforation depth and subtracting it from the bottom perforation depth. While this calculated interval is in open through the entire interval, it is common practice within the fluid extraction industry to define a "communication interval" or completion interval based on top and base perforations as these perforations establish communication with the geologic formation and allow fluid movement from the pore space into the pumped wellbore (Fetter 2001). Where no such openings are present, no fluid flow or communication may occur due to cement holding the casing in place above and below the openings (perforations) within the subsurface. To further

understand this concept, the reader is encouraged to refer to Figure 4 and take note of the lines running perpendicular to the well casing. These are completion or communication intervals.

3.2.11. Annual Subsidence Rate

Annual land subsidence, that behaves as the independent variable in the spatial regression model, comes from a raster mosaic dataset that may be downloaded from DWR. Via the DWR GIS image server, the designated time frame of this study was selected (2015-2021). This raster mosaic was then added to ArcGIS Pro in which the annual land subsidence rate was saved as an individual raster (.TIFF) for each respective year. Each of these annual rasters establishes the ground truth to compare the GWR results against.

To set the designated land subsidence value on a per well basis, the Extract Values to Points (Spatial Analyst) Geoprocessing tool was utilized in ArcGIS Pro. This process was repeated for each annual raster within the designated time frame of this study (2015-2021). This process was used to generate a pointset of wells with both groundwater level depths (bgs) and associated land subsidence from the SAR raster within the respective year.

3.2.12. Table of Variables

All datasets and associated layers listed in Table 4 are available from the starting year listed in the table and are current through the end of 2022 when this research was conducted, and this study was written.

Table 4 All datasets and associated layers listed in Table 4 are available from the starting year listed in the table and are current through the end of 2022 when this research was conducted, and this study was written. Table 4 outlines each independent variable used for viability in assessing spatial patterns of land subsidence in this study. As already established, these variables stem

from previous research, as well as professional geological experience. Each predictor variable holds a level of importance, but not all have been assessed together.

All datasets and associated layers listed in Table 4 are available from the starting year listed in the table and are current through the end of 2022 when this research was conducted, and this study was written.

Table 4. Ten independent variables used to assess land subsidence

Dataset	Variable	Originated From	Source	Starting Year
Well Completion Reports	Well depth	Well shapefile	DWR	1906
	Groundwater depth	Well shapefile	DWR	
	Annual average water level change*+	Well shapefile	DWR	
	Length of well completion+	Well shapefile	DWR	
Corcoran Clay Layers	Upper Tulare well completion+	Corcoran Clay layers	USGS	2009
	Lower Tulare well completion+	Corcoran Clay layers	USGS	
	Corcoran Clay thickness	Corcoran Clay layers	USGS	
	Corcoran Clay depth	Corcoran Clay layers	USGS	
Percent Coarse-to-Fine Grained Sediment	% fine-grained material	Well log ASCII	USGS	2009
	% coarse-grained material*	Well log ASCII	USGS	
SAR Mosaic	Annual subsidence rate	SAR rasters	DWR	2014

+ notates calculated value
* notates multicollinearity present

3.3. Exploratory Data Analysis (EDA)

Simple analyses were conducted to assess the state of the data and associated datasets. This includes the generation of histograms to assess statistical distributions of the data. The basics of exploratory data analyses to define normality of the datasets, residual diagnoses, Q-Q plots, Additionally, Kolmogorov-Smirnov tests (K-S Test) was conducted to test for normality alongside the Jarque-Bera test. Values found to be close to zero fit a normal distribution (or close to one). Any value not found close to zero demonstrates a non-normal distribution of skewness in

the data. This was conducted in R with the use of the “tseries” library and the “jarque.bera.test” function (Cromwell et al. 1994). Similarly, when assessing spatial regression at a global scale ArcGIS Pro was used to yield another run of the Jarque-Bera statistic. More on this can be found in the spatial regression section of this chapter. Datasets that were found to not be normally distributed were to undergo a log transformation. This then allows data to be analyzed using Original Least Squares (OLS) when it comes to the next step involving exploratory spatial data analysis (ESDA) (Chi and Zhu 2019).

Scatter plots were also generated to establish relationships such as total depth (TD) of water producers to surveyed groundwater depth. Additionally, groundwater depth to land subsidence were also plotted in scatterplot form. This approach brought the aspect of traditional statistics into the analyses. While these tend to be aspatial, each scatterplot and histogram was used to define strong and weak linear relationships between and among variables. It may also be noted that while such approaches are commonplace among GIS practitioners, as well as among geologists and engineers, there are no known publications that should the strength of linear relationships among such variables in the San Joaquin Valley; thus, the generation of a scatter plot matrix brought great insight about each explanatory variable as it association with land subsidence in the San Joaquin Valley.

For each dataset, descriptive statistics such as mean, median, minimum, and maximum values were generated and assessed. The generation of descriptive statistics allowed further assessment of each variable’s group of prosperities (Frost 2020).

3.4. Exploratory Spatial Data Analysis (ESDA)

The next steps were to perform ESDA to examine the spatial distribution of each variable. This included looking for global and local outliers, finding global and local trends, and

the examination of local variation as it coincides with spatial autocorrelation. Keeping in mind that this step in the process was to focus on multivariate datasets, it was also designed to establish both the neighborhoods of the individual wells and to determine the effective global and local search radius for each neighborhood (Griffith 1987).

The establishment of the neighborhoods is based on conducting a Moran's I Test. This was done using the "ape" library in R (Bivand et al. 2013). From here, the neighborhoods could be established as could the spatial weights matrix to help to correct for over-parameterization. This process allows explanatory variables to be dropped to then focus on the dependence relations that exist among observations and the variable Y_i (i.e. land subsidence) (Chi and Zhu 2019).

3.5. Multiple Linear Regression

Multiple linear regression (MLR) using OLS is often used to assess relationships between two or more variables, making it a candidate for assessing prediction functionality of continuous variables such as of land subsidence. MLR generates a model of variables to assist in understanding and quantifying variable relationships (Mitchell and Griffin 2021). It was with the ability to predict and quantifying multiple variables on a global level that MLR was used to assess patterns of land subsidence.

MLR, as with many forms of linear regression, can yield many statistics that help in assessing variables' relationships with each other. Each step, function, and statistical test outlined here was repeated for each annual dataset from 2015 to 2021. The Breusch-Pagan test was used for assessing existence of homoscedasticity (or heteroscedasticity) in the data. That is the assessment of non-constant standard deviations among independent variables. This too was initially conducted in R via the "lm" function alongside the "bptest" function (Breusch and

Pagan 1979). As with the normality test, the test of heteroscedasticity was again assessed in ArcGIS Pro via the use of the General Linear Regression Geoprocessing tool via the Koenker (BP) Statistic. These statistical tests were conducted in both ArcGIS Pro and in R as the initial runs in R were to assess variables as part of the EDA process.

Residuals were assessed for spatial biases in the form of errors and spatial lag via the use of the "spatialreg" library in R (Bivand et al. 2021). This step helps to build a relation between the response variable within an areal unit (i.e. the well location) that consists of a weighted average of the response variables at neighborhood areal units (Anselin 1988).

This exercise yields Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC) comparisons of the models and explanatory variables. The variables that are the most impactful, based on AIC and BIC results, were assessed and low performing explanatory variables were removed from consideration in the spatial regression model. The explanatory variables that were strong performs and helped to keep a low AIC were retained for use in global and local spatial regression models. Variables that were retained after assessment of the AIC values included total well depth, top perforation depth, well completion length, upper vs. lower Tulare completions, annual groundwater level, change in groundwater level, percent fine-grained material, percent-coarse grained material, Corcoran Clay thickness, depth to Corcoran Clay. Such results suggested that base perforation depth remain out of the spatial analyses for better model performance.

The Generalized Linear Regression (in the Spatial Statistics toolset) Tool was utilized in ArcGIS Pro. This tool can be used to assess continuous data relationships between two or more data attributes. Generalized Linear Regression (GLR) tool tends to take on the form of OLS when normally distributed, continuous datasets are assessed (Esri 2022). It must be noted that

GLR is inherently an aspatial regression tool as it takes on the form of a global regression model (Nelder and Wedderburn 1972). Additionally, GLR in the Esri toolbox is different from GLM in traditional statistics. For this study, GLR was used as it is an ArcGIS Pro tool that behaves like OLS despite the confusion in Esri's choice to name the tool in the contrary.

The explanatory variables that did not demonstrate multicollinearity were entered as the explanatory variables for MLR. Of course, subsidence was the dependent variable. In ArcGIS Pro, MLR is powerful in that it simultaneously produces measures of model performance (AIC_C), measure of goodness of fit (R^2), overall model statistical significance (Joint F-Statistic and Joint Wald Statistic as well as the Koenker (BP) statistic), and indicators of normality for residuals (Jarque-Bera statistic). Each of these was recorded for comparison with the local regression model.

Along with the above statistical measures, features classes of the residuals and standardized residuals were produced. These were subsequently mapped in ArcGIS Pro and assessed alongside model performance diagnostics.

3.6. Geographically Weighted Regression

The annual GWR models created in this study were generated in ArcGIS Pro. The inputs for the model included all explanatory variables that demonstrated no multicollinearity in the ESDA stage. The annual groundwater change was to be expected to one of the explanatory variables based on results from previous studies, albeit at a different location on the globe (Chu et al. 2018; Ali et al. 2020; Chu et al. 2021).

Within ArcGIS Pro, three different regression models may be considered. For this study Continuous (Gaussian) Model Type was chosen. This model type is effective when taking into consideration a range of values such as depth to groundwater level or even well total depth. As

determined by the name, the data used in this model type need to be normally distributed and data belong to a continuous data type (Mitchell and Griffin 2021).

Using the Moran's I Test results, the Neighborhood Type input was set to "Distance Band". As the effective distance for each neighborhood has been determined through ESDA efforts, this justified the use of "Distance Band" when establishing the GWR model

Finally, the local weighting scheme was set to Gaussian as it assigns a weight of one to the regression feature and the surrounding neighbors, *i* and *j* respectively (see Equation 4 from the previous sections within Chapter 3 for a quick refresher as to why this is established in this manner) (Fotheringham et al. 2022).

This same process was conducted for the designated timeframe of this study (2015-2021). The most influential variables were be maintained, and all parameters were kept consistent for every annual dataset.

The report output for GWR in ArcGIS Pro yields model diagnostics comparable to OLS. These include model diagnostics pertinent to measures of model performance (AIC_C), measure of goodness of fit (R^2), Sigma-Squared, Sigma-Squared MLE, and effective degrees of freedom. Each of these was recorded for comparison with the global regression model. Of note, each of the GWR results in the form of AIC, R^2, and adjusted R^2 are averaged as there are numerous results for the study area.

3.7. Model Comparison

After having constructed global and local regression models for predicting land subsidence, the performance of each model was assessed for accuracy. Each global regression model's AIC_C value was compared to each local regression AIC_C. The same was done for each model's coefficient of determination (R^2 and adjusted R^2). This was done in a simple table

comparison of values but was also assessed through the mapping of each global regression model's R^2 values and residuals as well as each local regression model's R^2 values and residuals. It is thought that strong performing models for predicting land subsidence will agree with Chu et al. (2018)'s results of R^2 values greater than 0.3 and generally no greater than 0.94.

An additional step included using the predicted land subsidence values from the GWR results, as well as the predicted land subsidence values from the MLR model results to run interpolation of predicted land subsidence values with the use of Inverse Distance Weighting (IDW). IDW is an accepted interpolation method for temporally based groundwater drawdown mapping (Fetter 2001; Galloway and Burbey 2011; Ali et al. 2020; Chu et al. 2021). However, there is an opportunity to improve upon this interpolation method for land subsidence results. Unfortunately, such a study is not in the scope of this research so the accepted method of IDW was used. The parameters that were used for IDW mapping of subsidence include a smoothing factor or 0.5, and output cell size of 100m (328 ft). Again, this emulates the work of Chu et al. (2021), but it is also noted that this visual comparison has likely introduced a second layer of error into the assessment of model performance. After this simple raster math subtraction was conducted between the original SAR datasets and the global regression model predictions. The same was done between the original SAR datasets and the local regression model predictions. The resulting delta rasters were then taken back to EDA through the generation of a histogram of the delta values. This allows a good assessment of the summary statistics while also showing how these data are distributed and if there is an acceptable range of error which is +/- 0.02 ft as established by Chu et al. (2018).

While this study is not about interpolation methods, the thought was that following Chu et al. (2021) and Ali et al. (2020) in the use of IDW would in fact yield a raster that can be

visibly compared to the annual SAR data. This visual comparison is to “see” the impacts of the values predicted alongside the quantitative regression diagnostics model performance indicators.

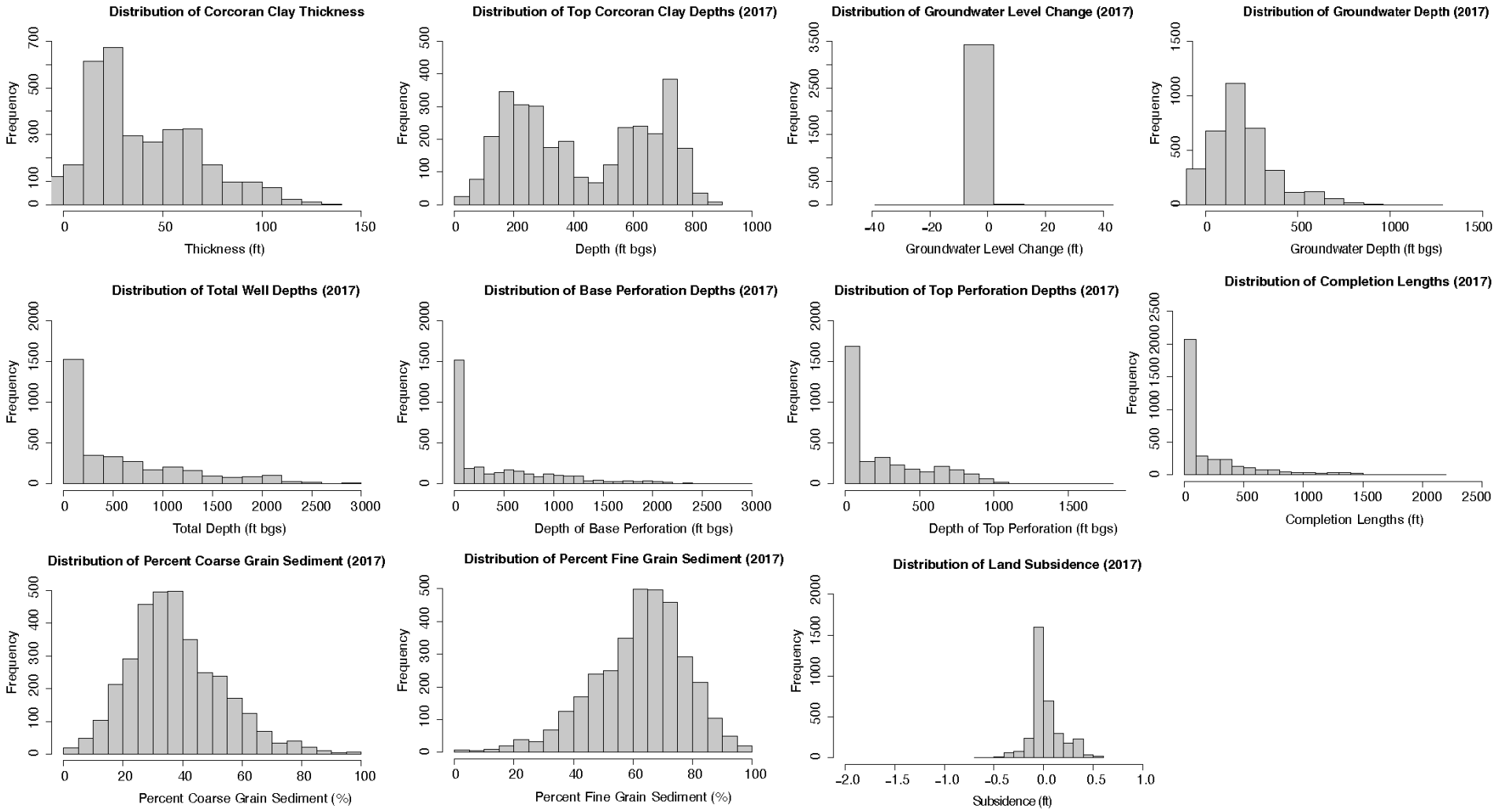
Chapter 4 Results

This chapter outlines patterns identified in geographically weighted regression (GWR) and statistical diagnostics of GWR spatial regression models. GWR accurately models land subsidence patterns based on key geological and engineering-based variables. GWR performs better than OLS models. Global regression techniques, such as OLS and spatially lagged models, were not as effective in assessing land subsidence patterns in the San Joaquin Valley. While such models have been successful at a global scale, such are not as impactful as local regression models like GWR. GWR further shows that proper predictor variables can have the greatest, positive impact when assessing land subsidence patterns.

4.1. Exploratory Data Analysis Distributions and Trends

The frequency curve of each predictor variable was assessed for similar values, with some being higher and some lower, for a normal distribution of the classic symmetrical bell curve. With this effort, values associated with each variable were determined to cluster around the center of the curve. This implies whether each predictor variable was normally distributed or not. Figure 8 shows distributions of variables associated with the 2017 dataset.

Figure 8. Histograms of 2017 explanatory variables and dependent variable



Assessing each distribution of both explanatory variables and the predictor variable (land subsidence), led to the use of the Jarque-Bera test to test the normality. Recall that for p-values of the test > 0.05 implies a normal distribution of the data. Similarly, a p-value of the test < 0.05 implies data with a distribution that is not normal.

Table 5 outlines the Jarque-Bera test p-values and implied level of statistical significance for the 2017 dataset.

Table 5. The Jarque-Bera test results for candidate variables

Dataset	P-value
Land Subsidence	0.7254
% Fine-grained Material	0.1941
% Coarse-grained Material	0.6681
Well Completion Length	0.382
Top Perforation Depth	0.5661
Base Perforation Depth	0.8378
Well Total Depth	0.8375
2017 Groundwater Level	0.4921
Groundwater Level Change	0.1832
Upper vs. Lower Tulare	0.6922
Depth to Corcoran Clay	0.8847
Corcoran Clay Thickness	0.06922

Such an aspatial assessment of these data show no statistically significant variable is present. Therefore, the model residuals were normally distributed or significantly biased. This also implies no data transformation is required among the associated variables (i.e. there is no need for a log transformation). For the 2017 dataset, as shown in Table 5 and exhibited in Figure 8, the Corcoran Clay Thickness comes close to being statistically significant. This makes sense due to the variability in thickness of this aquitard throughout the San Joaquin Valley. Several other geological factors, that show spatial patterns of their own, may also influence this variable

as outline in Chapter 3 (e.g. erosion and non-deposition). However, this variable remains normally distributed with no need for a data transformation. Except for these two variables, all explanatory and dependent variables are normally distributed with p-values > 0.05. The average of p-values for all variables from 2015 to 2021 is 0.474 further demonstrating a normal distribution of each variables' underlying values.

When assessing spatial regression models, the dependent variable, in this study land subsidence, also needs to be of a Gaussian distribution. Such a distribution is best when modeling spatial phenomena and patterns at global (OLS) and local (GWR) scales. Figure 9 is a histogram of the subsidence variable from the 2017 dataset that displays a normal distribution.

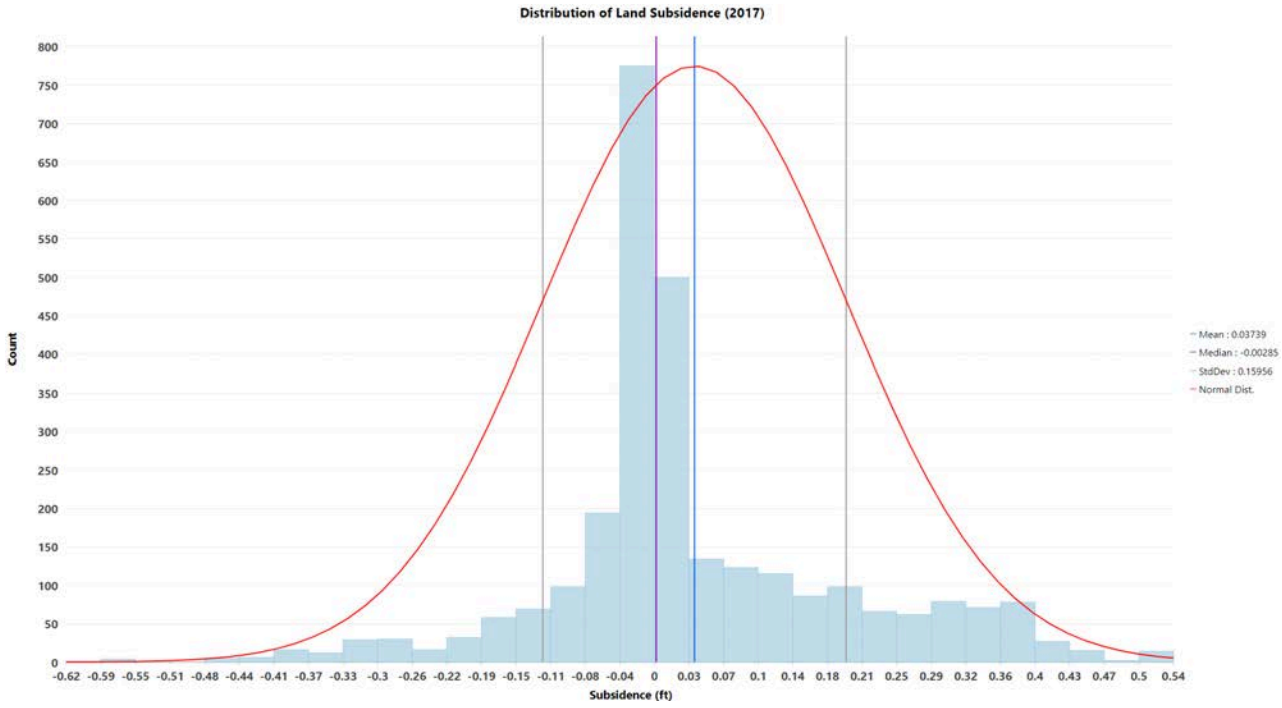


Figure 9. Histogram of land subsidence from 2017

When displayed in a scatter plot matrix, the linear relationship among each combination of both dependent and independent variables is exhibited. Complementing Table 5, Figure 10 outlines the R^2 values of each variable comparison in blue boxes on the upper right half of the figure. R^2 are posted inside these blue boxes. One may note how well total depth and well completion length have a high R^2 value (0.59). Yet land subsidence (this study's dependent variable) and well total depth (an independent variable of this study) have a low coefficient of determination (0.19) showing a large variance between these two variables.

Among the scatterplots in Figure 10, it is easy to identify that there is not a direct and single linear fit that can easily represent all the variables, as a two-way combination, at a global scale. However, it may be noted that all of the best-fit trends appear to be positive, and no variable combinations exhibit a negative relationship.

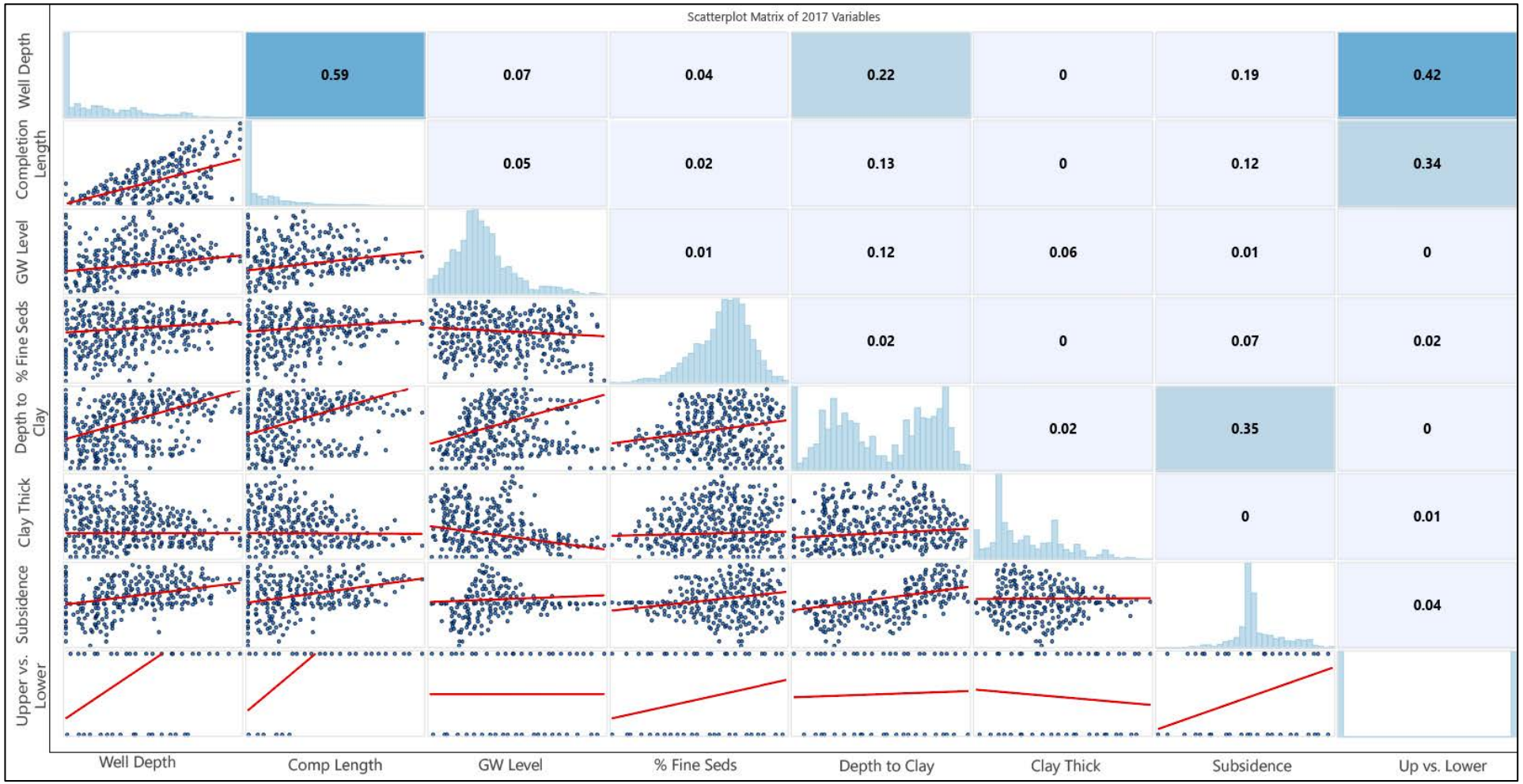
Of the histograms, most notably are those associated with the depth to the Corcoran Clay, the upper vs. lower Tulare, and the Corcoran Clay thickness variables. Both histograms associated with the Corcoran Clay appear to exhibit bimodal distributions. However, and as previously mentioned, traditional statistics have shown that the Corcoran Clay thickness is not statistically significant. And even though the upper vs. lower Tulare variable operates as a categorical dataset, it still exhibits a normal distribution.

The coefficients of determination, or R^2, that are in the blue shaded boxes Figure 10 show the strength of relationships for two-way variable combinations. As previously mentioned, there are variables that demonstrate a good correlation when combined, but there are also variables that demonstrate very weak measurements of variance for the dependent variable as it is explained by the independent variable. A good example of this is the R^2 value of 0 between the

depth to the Corcoran Clay and the total well depth. In short, neither of these variables can be used to predict the other.

Keeping these relations and distributions in mind, the next few sections will outline how these variables are spatially distributed and how these explanatory variables are used to predict patterns of land subsidence.

Figure 10. Scatterplot matrix of variables, distributions and R2 values, 2017



4.2. Local Cluster and Outlier Trends

Table 6 shows results from the Moran's I for spatial autocorrelation and the resulting search distances for the overall clustering of groundwater wells. Keep in mind that the search distances yielded through this process were utilized for each annual dataset when GWR was assessed. The incremental z-score graph yielded the values found in Table 6 for suggested search distances.

Table 6. Moran's I suggested search band distances

Year	Search Distance (ft)
2015	27004.2384
2016	27004.2384
2017	38670.74796
2018	27004.2384
2019	33994.90988
2020	84566.24516
2021	84564.56668

With these search distances, a larger cluster of high subsidence values were identified by the Anselin Local Moran's I statistic for almost all wells located on the west side of the San Joaquin Valley in 2017. These values are displayed in light red in Figure 11 and are likely associated with high values of aquifer recharge and inflation. However, a rim of low-high outliers is present around this region, which likely signifies groundwater extraction wells associated with agricultural crop irrigation. The light blue values indicate low-low clusters by which there is a low clustering of wells exhibiting small amounts of subsidence. Grey values to the northwest of Merced and to the west of Modesto are well locations that had no statistically significant land subsidence or inflation in 2017.

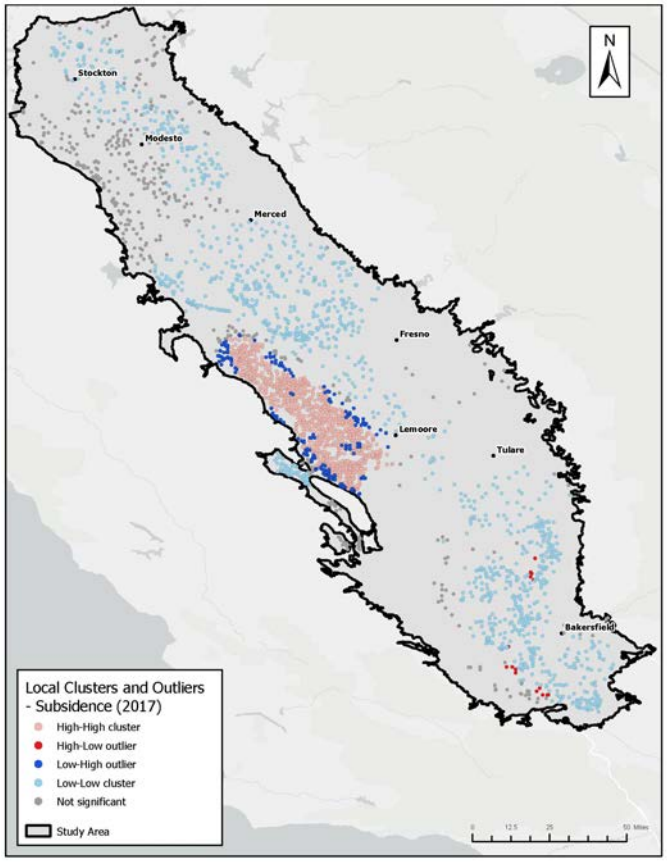


Figure 11. Map of land subsidence clusters and outliers, 2017

High-high clusters made up 32% of the subsidence clusters (912 out of 2,815) for the 2017 dataset. Meanwhile, low-low clusters made up 44% of the subsidence clusters (1,246 of 2,815) for the 2017 dataset. 16% of the subsidence clusters were found to not be statistically significant (444 of 2,815) for the 2017 dataset.

As Figure 11 indicates, there are small amounts of high-low and low-high outliers found throughout the valley. Only 1% of all the well locations were outliers with high values surrounded by low values (15 out of 2,815). Most of these low-high outliers are found in the southern portion of the valley and are mostly west of Bakersfield. In assessing these locations, it may be noted that these values are not related to oil or gas extraction which that region is known

for. Rather, like the rest of the data in this study, are related to high volumes of groundwater extraction for crop irrigation that led to higher rates of subsidence in 2017. The remaining high-low outliers were previously identified in areas showing inflation due to aquifer recharge. High-low values accounted for 7% of the clusters and outliers from the 2017 dataset (198 out of 2,815). Figure 12 explores the breakdown of outliers and clusters further. This Moran's scatterplot further demonstrates the clustering of high-high values as one would expect to see what looking at localization of land subsidence due to groundwater extraction in an agriculturally rich valley.

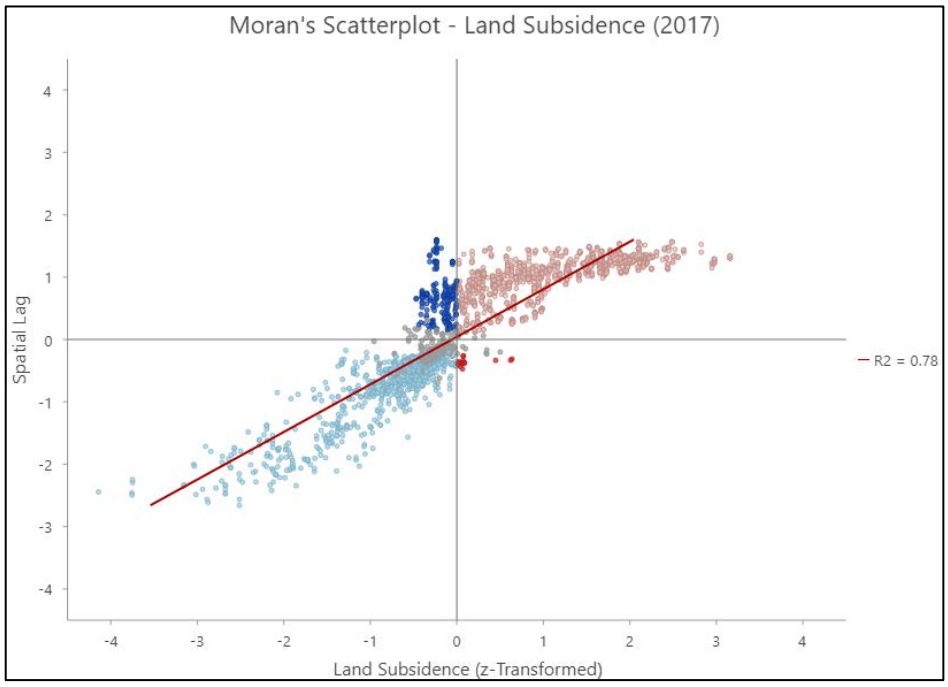


Figure 12. Moran's scatterplot of land subsidence values, 2017

Figure 12 further illustrates the point that neighboring water wells exist in areas of similarly high or low rates of land subsidence. One may note that the low-low both figures demonstrate clusters of subsidence throughout the San Joaquin Valley. One must further keep in

mind that low values indicate rates of elevation drop or subsidence. Something that would be expected in large portions of the valley. Such low-low clustering that is observed does in fact show subsidence occurring throughout the study area. The next question is at what level of variability exists among these subsidence clusters?

Visual spatial patterns of land subsidence trends and relationships are explored in a more quantitative way among the previously outlined global regression analyses. Further refining of such spatial pattern may be found in the next section pertaining to local regression analyses.

4.3. Global Relationship Trends

Global regression was assessed without the inclusion of spatially lagged explanatory variables. This was done through OLS and through the creation of a single, global model of each variable being used to predict land subsidence. The OLS for 2017 explanatory variables yielded a single variable that is not considered to be a statistically significant variable. That variable was associate with groundwater level change in 2017 that has a p-value of 0.3095. All the other nine explanatory variables were found to be statistically significant. Table 7 presents summary results for each explanatory variable and its measure of strength when regressed with land subsidence. Table 7 is a more in-depth version of Figure 10 that shows other simple regression statistics.

Table 7. Summary of two-variable regression results, 2017

Variable	P-Value	R2	Adj R2	Std. Error	t-Statistic	f-Statistic
Groundwater Level Change	0.3095	0.0003	9.61E-06	0.12842	1.016	1.033
Groundwater Level 2017	2.348E-06	0.006457	0.006168	17.494	4.729	22.6
% Fine Grain Sediment	2.2E-16	0.0601	0.05983	1.6924	14.82	219.6
% Coarse Grain Sediment	2.2E-16	0.061	0.05983	1.6924	-14.82	219.6
Completion Length	2.2E-16	0.1286	0.1284	34.62	22.54	508

Top Perforation Depth	2.2E-16	0.1757	0.1755	30.771	48.18	733.7
Base Perforation Depth	2.2E-16	0.1832	0.1829	58.094	27.78	771.6
Well Total Depth	2.2E-16	0.2009	0.2007	65.11	29.41	865.2
Upper vs Lower Tulare	2.2E-16	0.04851	0.04823	0.054602	13.14	172.8
Corcoran Clay Thickness	0.4484	0.0002043	-0.0001511	0.5233	75.943	0.5749
Depth to Top Corcoran Clay	2.2E-16	0.3469	0.3467	3.894	99.97	1494

As Table 7 shows, the t-statistic yields a significant change in magnitude, mostly in the positive direction, but with one variable, percent coarse-grained material in the negative direction. The t-values are significant as each variable's absolute value is higher than 1.96 ($|t| \geq 1.96$).

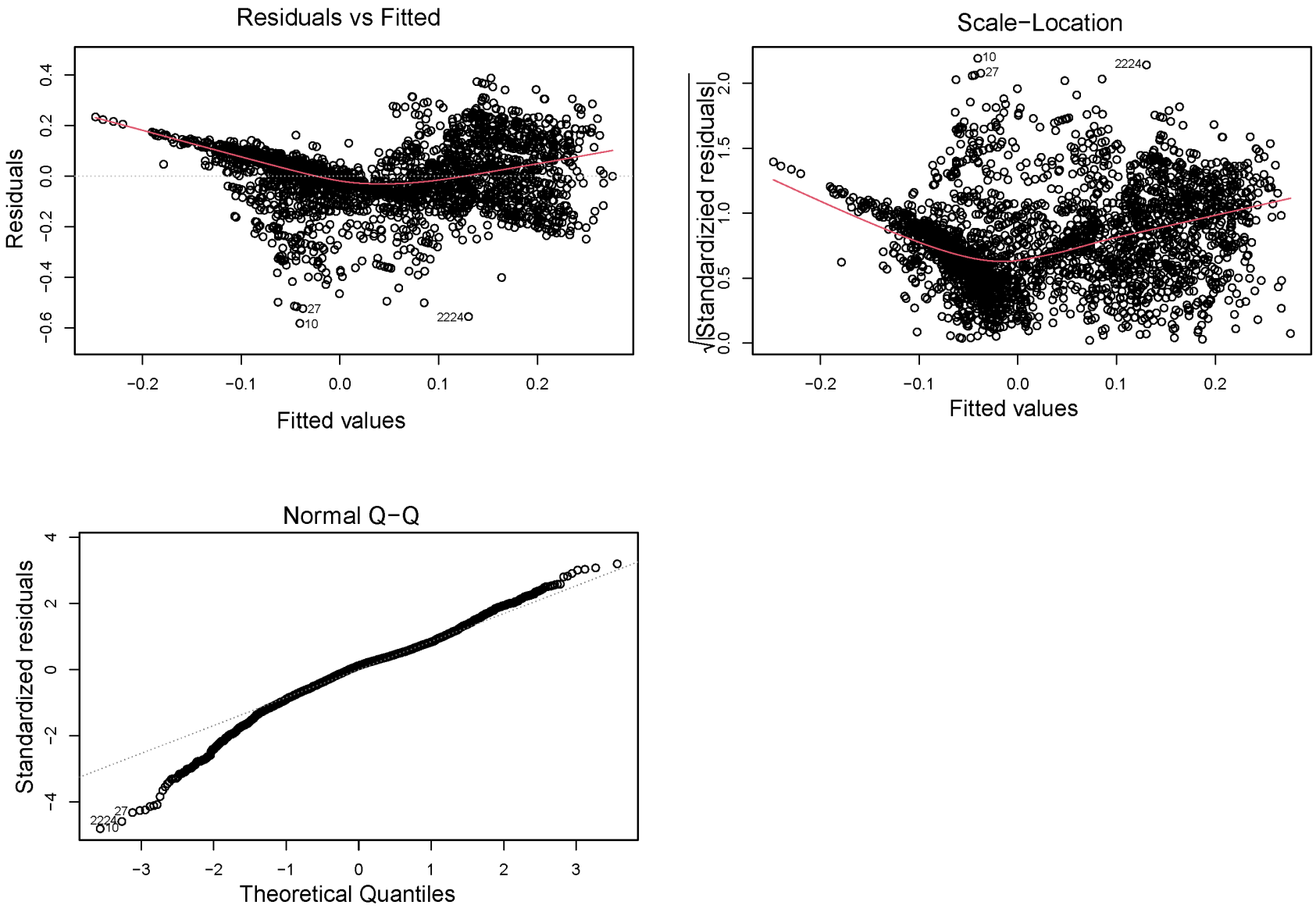
Table 7 goes on to show that f-statistics associated with each variable tend to be large, again apart from the groundwater level change variable. One must keep in mind that the larger the number, the larger the dispersion of data from the mean. The variables top perforation depth, base perforation depth, and total well depth display the largest ration of variance, or dispersion, than any of the other variables.

The recorded standard errors show large values (> 0.9). The variables that are an exception to large standard error include all values related to well depths (e.g. top perforation depth, base perforation depth, and total well depth) and groundwater level depths for 2017. It should be noted that that once again the variable associated with groundwater level change has a lower standard error.

Finally, R^2 values, as well as the adjusted R^2 values, were relatively low for each variable in the OLS model. The highest recorded R^2 value was 0.2009 in the 2017 dataset. This estimate of movement between dependent and independent variable was associated with total well depth.

Figure 13 exhibits plotted values as they pertain to residuals vs fitted, Normal Q-Q plot, and standardized residuals. One may note the deviation from a linear fit on all three graphs.

Figure 13. Residual vs fitted and Normal Q-Q plot



The residuals vs. fitted values plot shows no direct linear correlation among variables. Assessing the spread of the data on the plot shows that there is some homoscedasticity for residuals from 0.0 to 0.2. These also follow what starts as a linear trend for the matching fitted values from -0.2 to -0.1. From there, these data exhibit high levels of heteroscedasticity further implying a non-linear model at the global scale. This is further emphasized with the red best-fit line being non-linear.

The normal Q-Q plot are assessing a possible normal distribution of the standardized residuals over the theoretical quantiles. As values on this plot in the lower left do not show sets of quantiles aligning with the dashed best-fit line, it may be said that these quantiles do not come from the same distribution—namely, they do not come from a normal distribution at least when assessed at the global scale. There is a short time in which the quantiles do follow a linear trend. This is noted on the x-axis between -1 to ~1 where the compared quantiles plot on the dashed linear fit line. It also appears that around 95% of the data lie below 2.80.

When assessing the standardized residual vs fitted values plot, the measure of the strength of the difference between observed and expected values of a linear model is shown. In this case, much like with the residuals vs. fitted values, there is no linear relationship exhibited. In fact, the standardized residuals show a large separation between the linear model's observed and expected values. Once again, this demonstrates that a large, global model may not be the best approach to prediction land subsidence.

4.3.1. Multiple Linear Regression (MLR)

Global linear regression was further assessed via the use of Generalized Liner Regression (GLR) in ArcGIS Pro. One must keep in mind that MLR is intended to model the dependent variable based on a grouping of independent or explanatory variables (Mitchell and Griffin 2021). While this approach follows an attempt to fit a continuous model (OLS), it considers all

explanatory variables in a single snapshot as opposed to unique, individual variable comparisons (i.e. lone explanatory variable compared to dependent variable). Figure 14 exhibits the standardized regression residuals of MLR for the 2017 dataset. Note that the dark red values signify higher negative values from the standard deviation, while the darker blue signifies higher positive values from the standard deviation. Figure 14 further emphasizes where large error residuals reside. In the case of both maps, west of Lemoore and south of Merced tend to have largest residual errors.

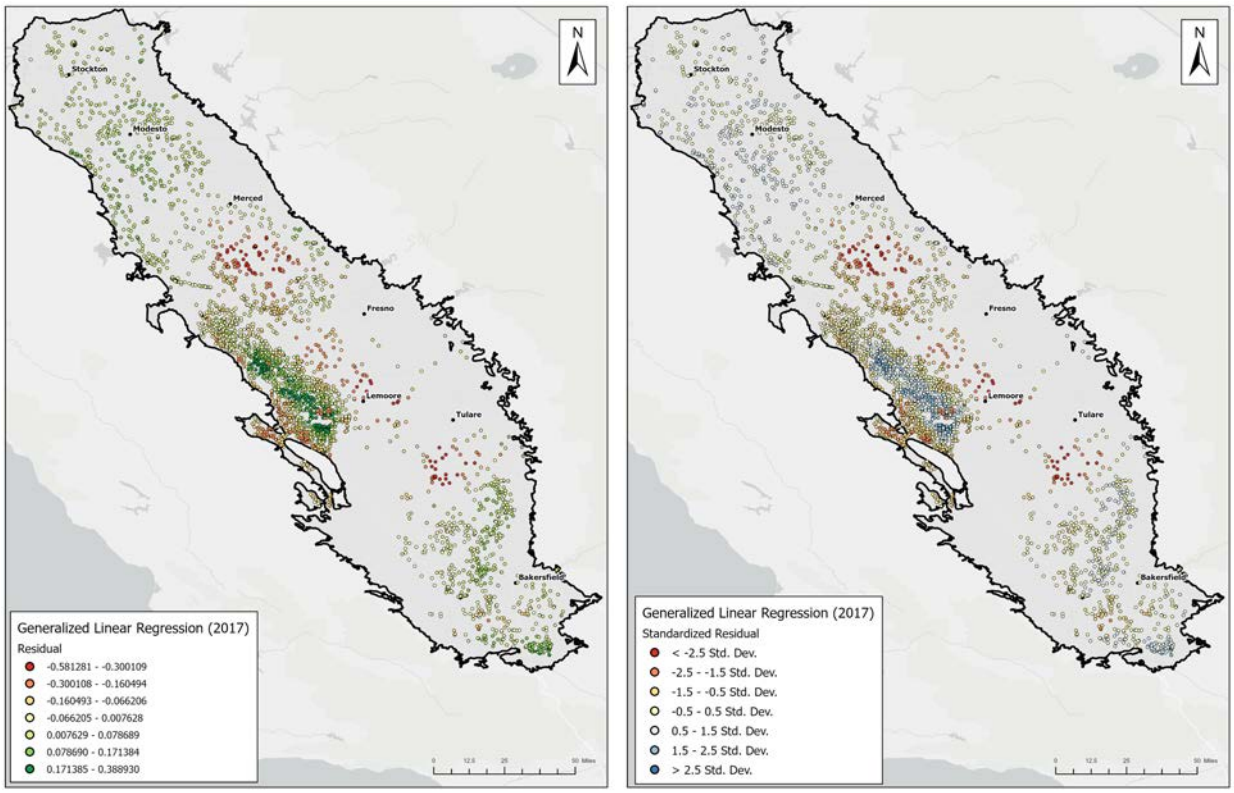


Figure 14. Map of 2017 MLR residuals and standardized residuals

A summary of OLS diagnostics results for 2017 is provided in Table 8. This table includes the coefficient, robust SE, robust t, robust Pr, and variance inflation factor (VIF). As previously mentioned, a VIF >5.0 indicates multicollinearity. For this reason, the annual groundwater change variable was removed from the analysis as all but one year demonstrated

multicollinearity with the dependent variable of land subsidence. OLS diagnostic results for all years (2015-2021) may be found in Table 8.

Table 8. Summary of OLS diagnostics, 2017

Variable	Coefficient	Robust SE	Robust t	Robust Pr	VIF
2017 Intercept	-0.160512	0.010603	-15.138003	0.000000*	--------
Total Well Depth	0.000024	0.000007	3.584269	0.000358*	3.726871
Well Completion Length	0.000022	0.000011	2.033164	0.042121*	2.557095
2017 Groundwater Level	-0.000156	0.000015	-10.254633	0.000000*	1.355889
Depth to Corcoran Clay	0.001209	0.000152	7.967612	0.000000*	1.093304
Corcoran Clay Thickness	0.000375	0.000013	29.571774	0.000000*	1.664723
% Fine Grain Material	-0.000626	0.000077	-8.096861	0.000000*	1.152584
Up vs. Lower Tulare	0.021195	0.006339	3.343754	0.000854*	2.179722

As VIF is mentioned here, it should be noted that when the independent variables were entered into the GWR Geoprocessing tool in ArcGIS Pro, a multicollinearity error was thrown. Assessing each variable individually in the GWR tool, allowed groundwater level change, top perforation depth, and percent coarse-grained sediment to be identified as exhibiting multicollinearity. This was particularly interesting as all VIF measurements had up to this point had no indication of multicollinearity., and as is only identified with the annual groundwater change variable. The choice was made to move forward with the remaining explanatory variables of well completion length, annual groundwater level, percent fine-grained sediment, depth to Corcoran Clay, Corcoran Clay thickness, and upper vs. lower Tulare aquifer completion, each is reflected in the results that follow.

Closely tied to the VIF in Table 8, are the joint F-statistic, joint Wald statistic, Koenker (BP) statistic, and Jarque-Bera statistic. Table 9 exhibits each of these statistical results that pertain to OLS model diagnostics. The Jarque-Bera statistic have previously been addressed to assess multicollinearity among variables and assessment of normal vs. non-normal distributions.

Most notably are the degrees of freedom and the numbers of observations for each OLS assessment. As previously mentioned, the error degrees of freedom are independent pieces of information that are used in estimating coefficients. For precise coefficient estimates, especially in regression testing, one should have many degrees of freedom as to have more observations for each model term. This is desired in any regression model and allows for more availability to calculate the desired coefficients. Each year analyzed demonstrates large degrees of freedom and should have resulted in a valid calculation of coefficients. Again, note how the degrees of freedom in Table 9 greatly exceed the number of observations.

Table 9. Summary of OLS Results for all annual datasets

Year	Number of Observations	Degrees of Freedom	Joint F-Statistic	Joint Wald Statistic	Koenker (BP) Statistic	Jarque-Bera Statistic
2015	2820	72812	122.841	1153.328	147.439	3797.019
2016	3286	73278	161.800	1171.711	183.536	5445.731
2017	2768	82759	260.612	2033.425	132.651	434.332
2018	3096	83087	72.035	641.122	125.710	13690.234
2019	1771	81762	18.876	166.498	106.255	2787.450
2020	1517	71509	124.501	448.850	219.968	1735.485
2021	1552	71544	167.955	682.999	199.268	1765.461

The joint Wald statistic for each year is well above zero and implies that the variables used in the MLR model are valid and should be included in the model. 2019 shows up as having the lowest joint Wald statistic (166.4987), but even with this value, it implies that all variables should remain in the model to assist with model fit.

The joint F-statistics for each year are valid and thus allow one to trust the Koenker (BP) statistic. However, as each of the F-statistic values are relatively high, each year's group average is more spread out than the variability of the data within each group. Here, the differences in the data averages likely reflect differences that exist at the population level in the data.

The Koenker (BP) statistic does not yield a value that is statistically significant (<0.05) for any of the years assessed. This would imply that there is stationarity present in the model and that the independent variables vary or fluctuate throughout the study area in relation to land subsidence. This further alludes to the fact that a local regression model is likely to help improve predictions.

Table 10 yields the end results of MLR and exhibits the AICc values from global regression.

Table 10. MLR AICc estimated prediction error values

Year	AICc
2015	-1616.5047
2016	-2337.7899
2017	-3802.4281
2018	-4387.8372
2019	-4455.4981
2020	-2437.5007
2021	-1894.7676

AICc values from the MLR annual models estimates the quality of each model, relative to each of the other model's results. Here it may be noted that the 2018 and 2019 models contain the lowest error and hence are the better of the seven models created through the MLR process for each annual dataset. These values are compared to the GWR results over the next couple of sections, but a direct AICc comparison is made in section 4.5.

4.4. Geographically Weighted Regression and Patterns of Land Subsidence

Local regression in the form of GWR included all previous independent variables assessed with MLR. With GWR, all measured values performed better than in the global regression model. From the previous section one may recall that the R^2 value for the 2017 dataset

was 0.430417 (adjust R^2 of 0.428766). The MLR model had an AIC_C of -3802.42091. As opposed to the GWR model that yielded an R^2 value of 0.8384 and an AIC_C of -6913.942 for the 2017 dataset. The negative AIC_C indicates a lower degree of information loss as opposed to a positive AIC_C (Baguley 2012). Table 11 exhibits GWR model diagnostics for each annual dataset (2015-2021).

Table 11. GWR model performance diagnostics by year

2015	
R^2	0.9044
AdjR^2	0.8871
AIC_C	-6765.8909
Sigma-Squared	0.0048
Sigma-Squared MLE	0.0041
Effective Degrees of Freedom	2388.0913

2016	
R^2	0.8956
AdjR^2	0.8773
AIC_C	-7128.8269
Sigma-Squared	0.0047
Sigma-Squared MLE	0.004
Effective Degrees of Freedom	2499.3111

2017	
R^2	0.8384
AdjR^2	0.8227
AIC_C	-6913.942
Sigma-Squared	0.0046
Sigma-Squared MLE	0.0042
Effective Degrees of Freedom	2521.6353

2018	
R^2	0.9106
AdjR^2	0.8954
AIC_C	-10594.457
Sigma-Squared	0.0017
Sigma-Squared MLE	0.0015
Effective Degrees of Freedom	2646.2194

2019	
R^2	0.8439
AdjR^2	0.814
AIC_C	-8208.716
Sigma-Squared	0.0008
Sigma-Squared MLE	0.0007
Effective Degrees of Freedom	1667.2951

2020	
R^2	0.8033
AdjR^2	0.7917
AIC_C	-4096.7792
Sigma-Squared	0.0038
Sigma-Squared MLE	0.0036
Effective Degrees of Freedom	1432.5391

2021	
R^2	0.7987

$AdjR^2$	0.7869
AIC_C	-3385.6421
Sigma-Squared	0.0064
Sigma-Squared MLE	0.0061
Effective Degrees of Freedom	1465.8185

4.4.1. Geographically Weighted Regression Coefficients

Mapped coefficients for independent variables for the 2017 dataset are part of the capstone of GWR as such maps offer insight to the factors that contribute to the outcome of the dependent variable. The key here is to assess how spatially consistent the relationship between dependent and explanatory variables are. Mapped coefficients show the distribution of variation; that is, how much variation is present and where, among these variables.

On assessing the distribution of completion length coefficients as shown in Figure 15, 2,122 locations exist with little variance in completion length when assessing patterns of land subsidence. This large clustering of locations is notably located in the Westside Subbasin. As previously mentioned, the majority of subsidence found int the 2017 SAR data may also be found in this subbasin. A large cluster of coefficients in this subbasin is not surprising and is likely attributed to well owner' desire to chase dropping groundwater levels with longer completions along the well bore. There is a degree of variance in the Modesto area as values change from the southside, through central portions of Modesto, and again on the north side of Modesto. It is also of note that little variation in completion length occurs south of Tulare and to the northwest of Bakersfield; however, there is a degree of variance between the Tule Subbasin and the Kern County Subbasin.

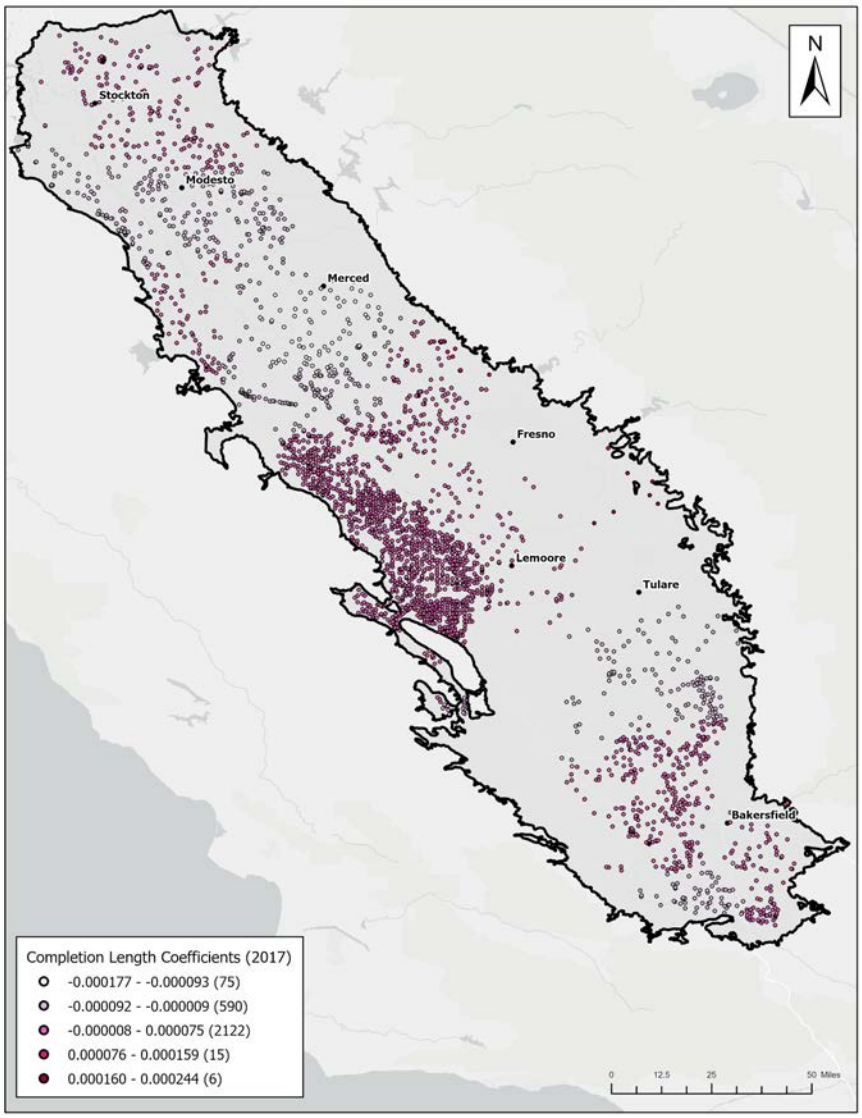


Figure 15. GWR completion length coefficients map, 2017

The Corcoran Clay thickness coefficients tend to follow spatial trends that are indicative of the lateral extent of the clay itself and are exhibited in Figure 16. Coefficients in the Westside Subbasin, west of Lemoore, contain negative coefficients implying that as the Corcoran Clay increases in thickness, subsidence decreases. However, one must keep in mind that negative SAR data tells us that there is inflation or uplift occurring in the area, so we must think the reverse of what regression coefficients would normally tell us. More simply stated, as the Corcoran Clay

decreases in thickness, land subsidence is in fact increasing. This concept fits with what Lees et al. (2021) discovered in their hydraulic head modeling, that is that the presence of the Corcoran clay does influence subsidence, but it does not mean that such a presence of fine-grained sediment should be ignored in unconsolidated sandy aquifers. These coefficients also align with Faunt et al. (2015)'s discovered that up to 30% of the overall subsidence occurs in the upper unconfined portion of the aquifer.

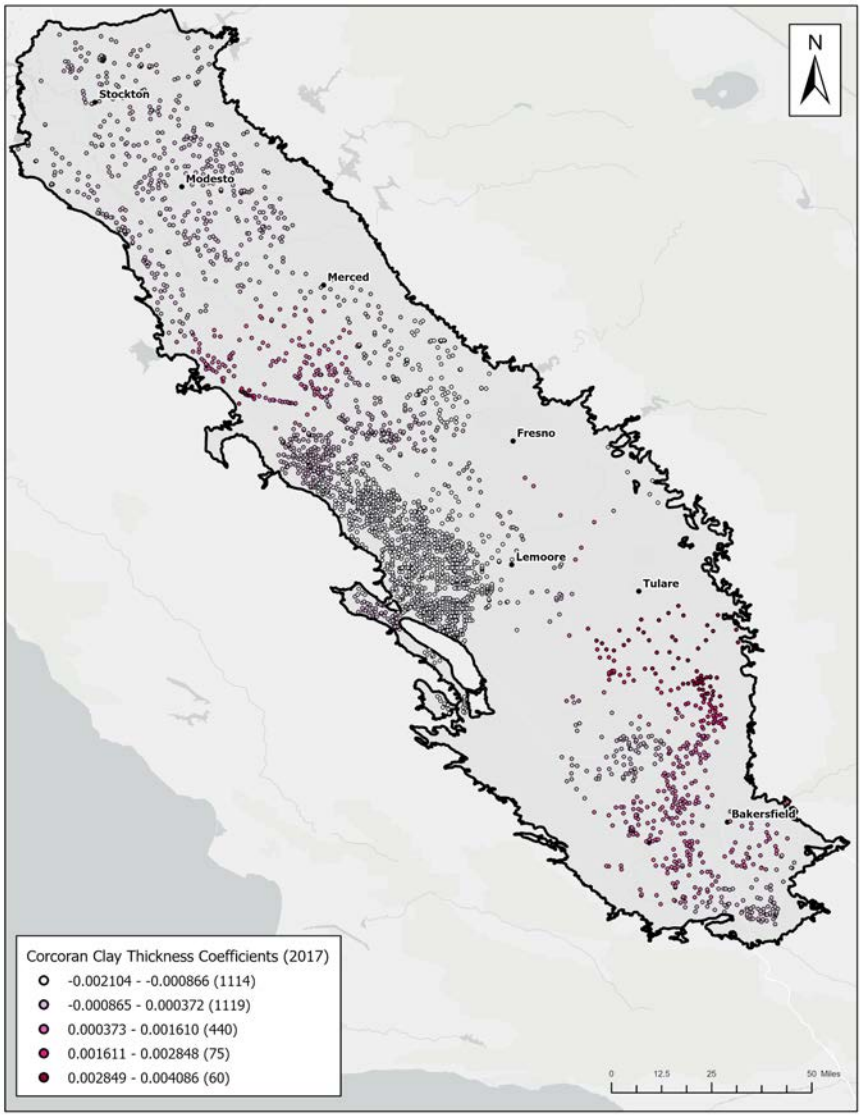


Figure 16. GWR Corcoran Clay thickness coefficients map, 2017

To further emphasize the importance of the Corcoran Clay, coefficients associated with the depth to the clay show similar clustering as the thickness of the clay. The big difference here is that around half of the coefficients exhibit a positive relationship so that as the depth to the Corcoran Clay, when it is present, increases, the amount of subsidence also increases. This is demonstrated in Figure 17. This positive relationship is most notable in the notorious Westside Subbasin. Negative coefficients exist around Modesto and even to the south and west of Bakersfield.

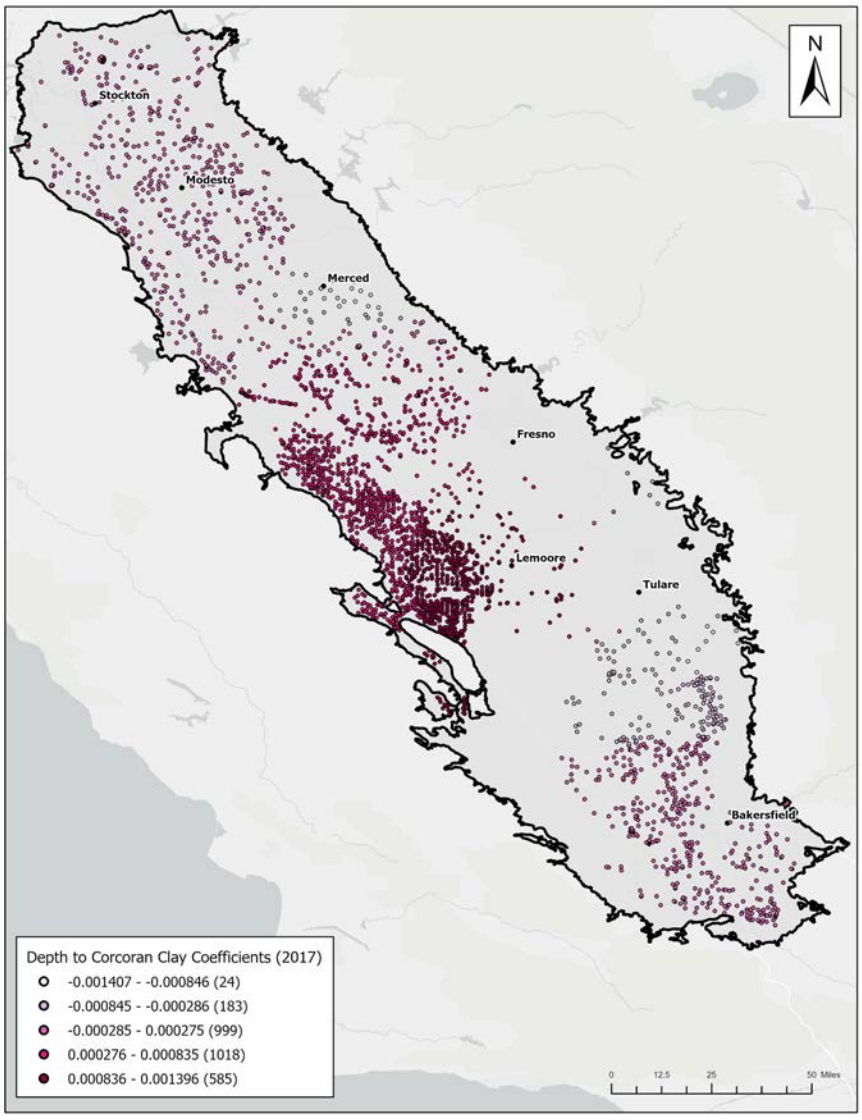


Figure 17. GWR depth to Corcoran Clay coefficients map, 2017

The importance of fine-grained sediment has been discussed throughout this study. Assessing the percentage of fine-grained sediment coefficients to land subsidence continues to demonstrate what the water industry and academia have discussed; that is that as the volume of fine-grained sediment increases, the amount of land subsidence will also increase. This is best shown in the Westside Subbasin where all coefficients are positive as exhibited in Figure 18. These coefficients change as one moves to the northeast where values between Fresno and Merced exhibit a negative relationship. The area around Modesto also shows negative coefficient values. The same can be said for Bakersfield. Some of this may be geologically related as the deposition of fine-grained sediments occurs in quite water, such as those of the base of the ancient Tulare Lake.

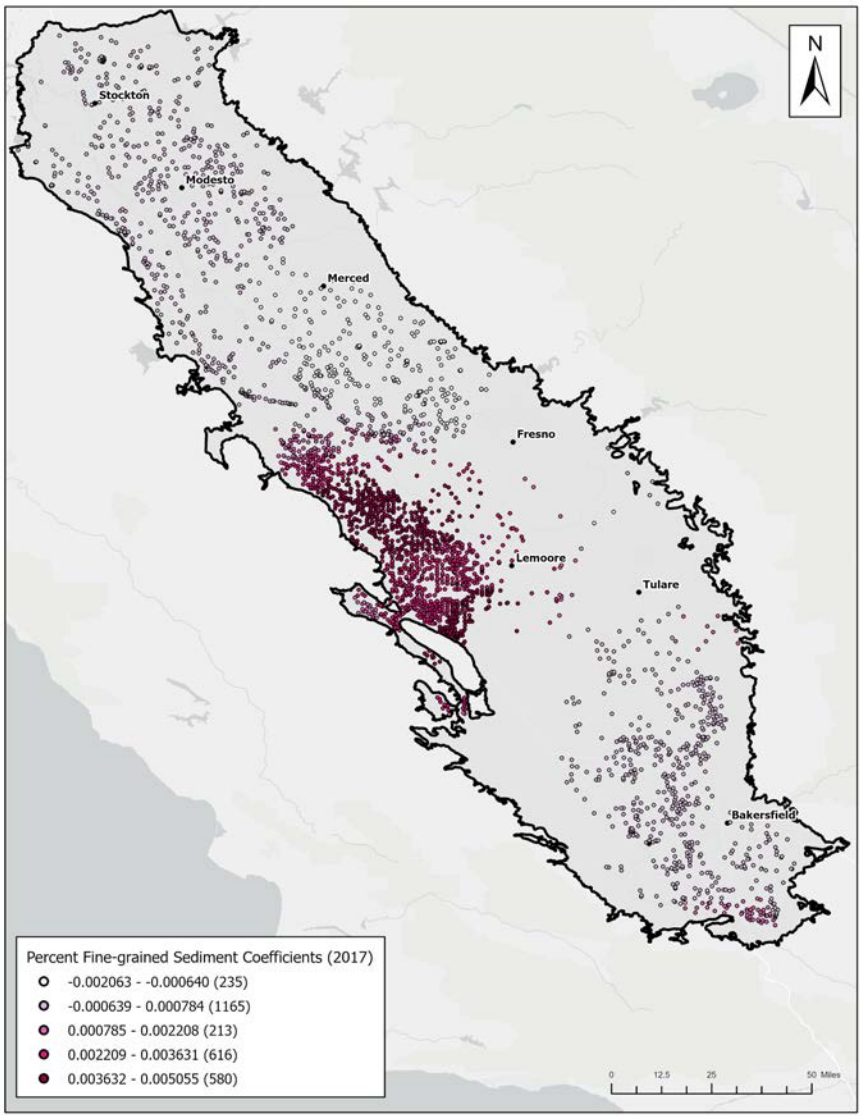


Figure 18. GWR percent fine-grained sediment coefficients map, 2017

For the groundwater level coefficients, it comes as no surprise that clustering continues in and around subbasins. In the Westside Subbasin, a negative relationship exists. This too is not a surprise as the mechanics of decreasing groundwater levels with increase the amount of land subsidence occurring. This again fits with the fact that groundwater drawdown is inextricably linked to land subsidence (Sneed 2018). Coefficient values around Modesto continue to show a negative relationship between groundwater level and land subsidence. This negative relationship

is also exhibited to the northwest of Bakersfield, but changes to a positive relationship to the southwest and is shown in Figure 19.

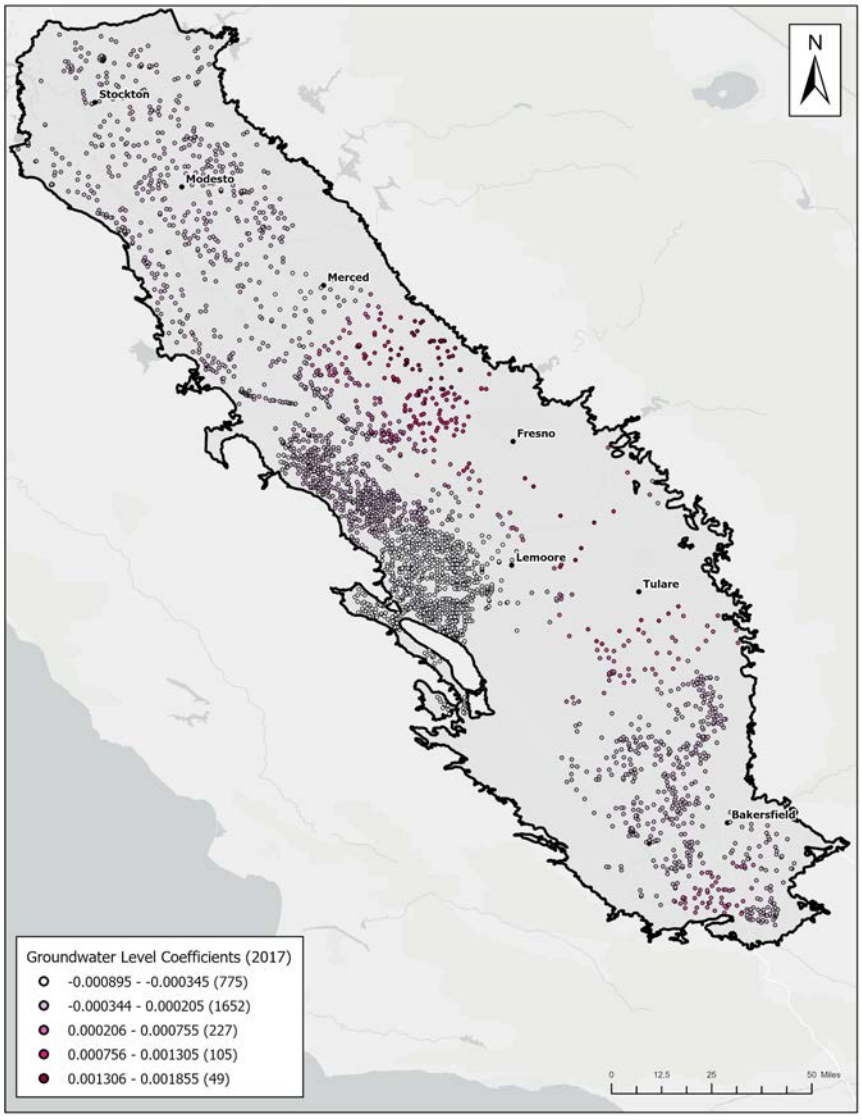


Figure 19. GWR groundwater level coefficients map, 2017

One of the interesting spatial distributions of coefficients comes in the form of the upper vs lower Tulare well completions that is shown in Figure 20.

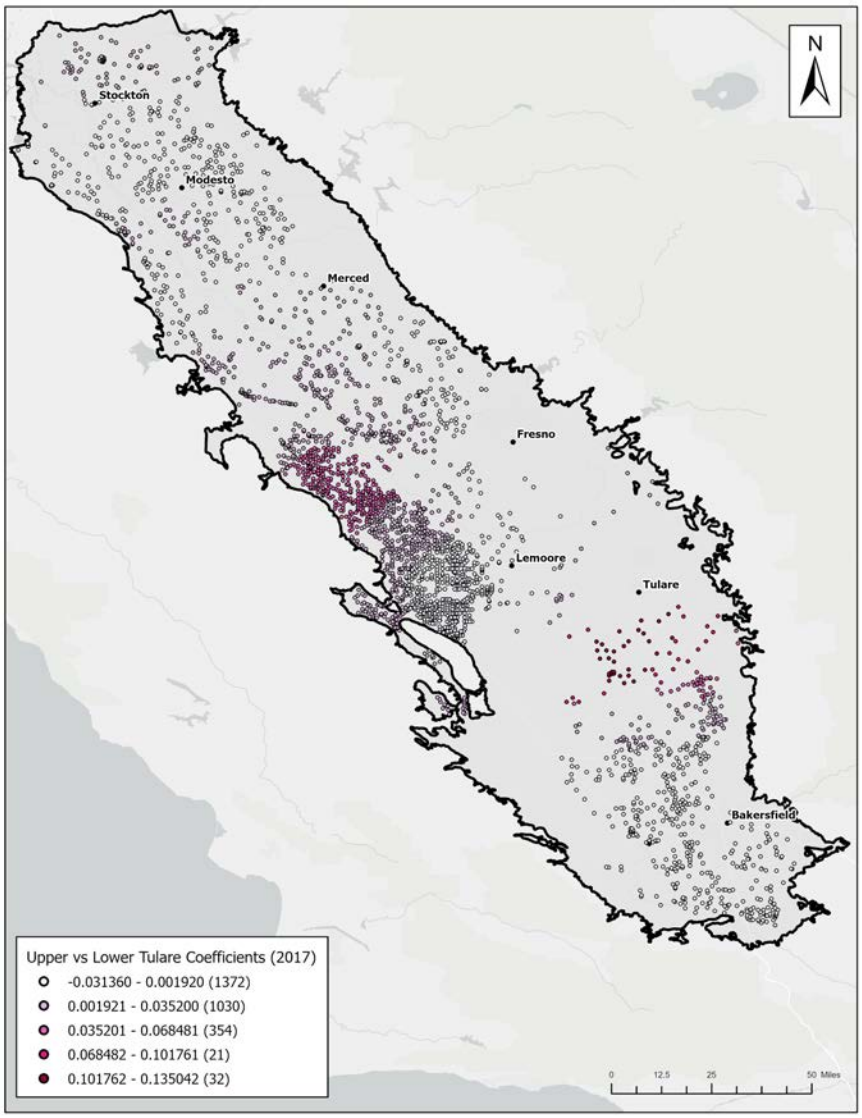


Figure 20. GWR upper vs. lower Tulare coefficients map, 2017

Figure 21 takes another perspective of the upper vs. lower Tulare coefficients through a zoomed in look at coefficient values in the Westside Subbasin.

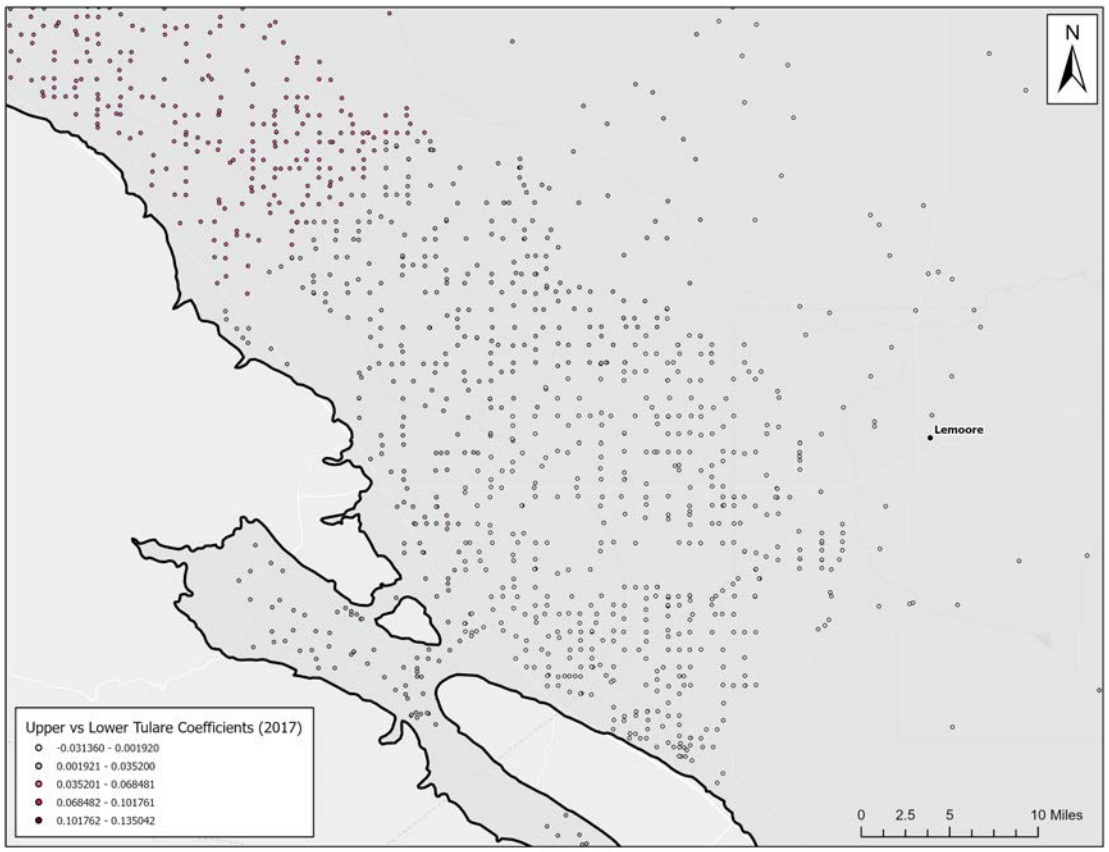


Figure 21. Distribution of upper vs lower Tulare completion coefficients on the westside

These coefficients translate from showing a negative relationship to a stronger and stronger positive relationship moving north. This is also interesting to see as the large variance among variables across this region further demonstrates why GWR is a great tool to help optimize land subsidence predictions.

Finally, the well depth coefficients also exhibit an interesting spatial distribution of relationships as is shown in Figure 22.

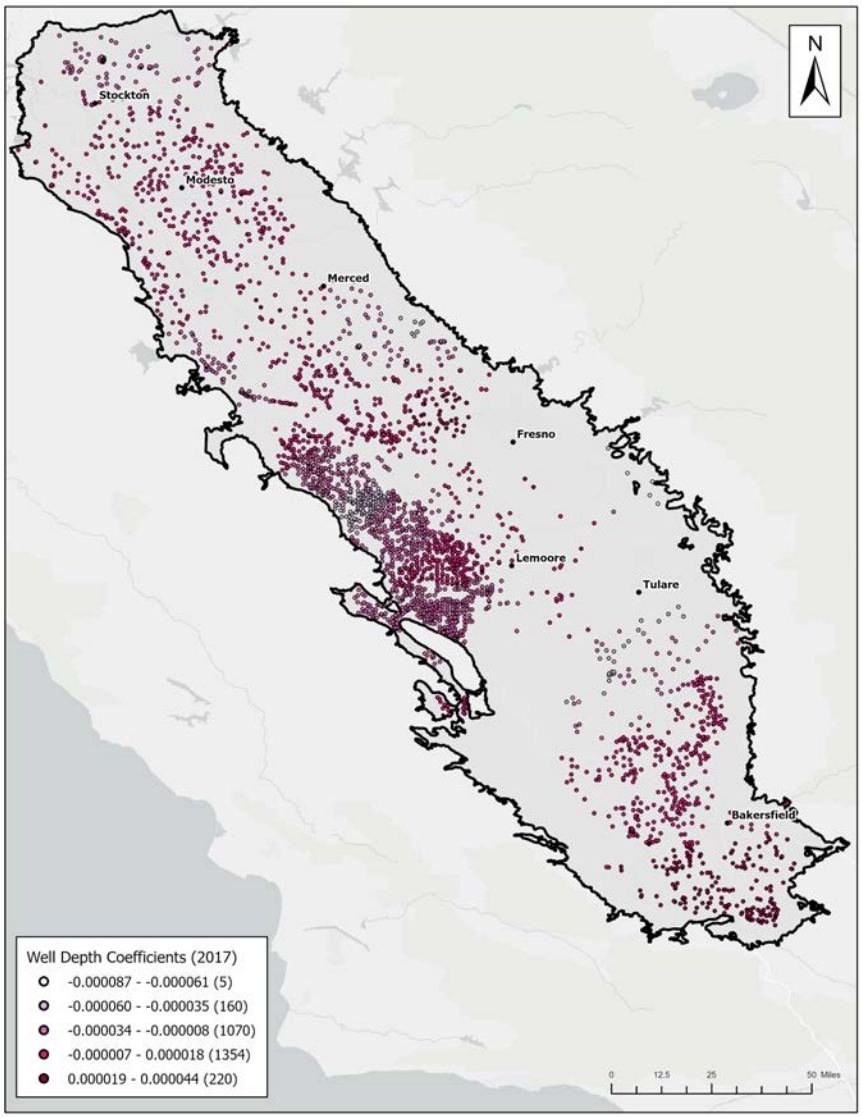


Figure 22. GWR well depth coefficients map, 2017

The majority of well depth coefficients demonstrate negative relationships. On the westside, among the Westside, Pleasant Valley, and Kettleman Plain Subbasins, such a relationship would imply that as wells increase in depth land subsidence decreases. Figure 23 is a zoomed in view along the westside of the study area and shows this spatial distribution of coefficients.

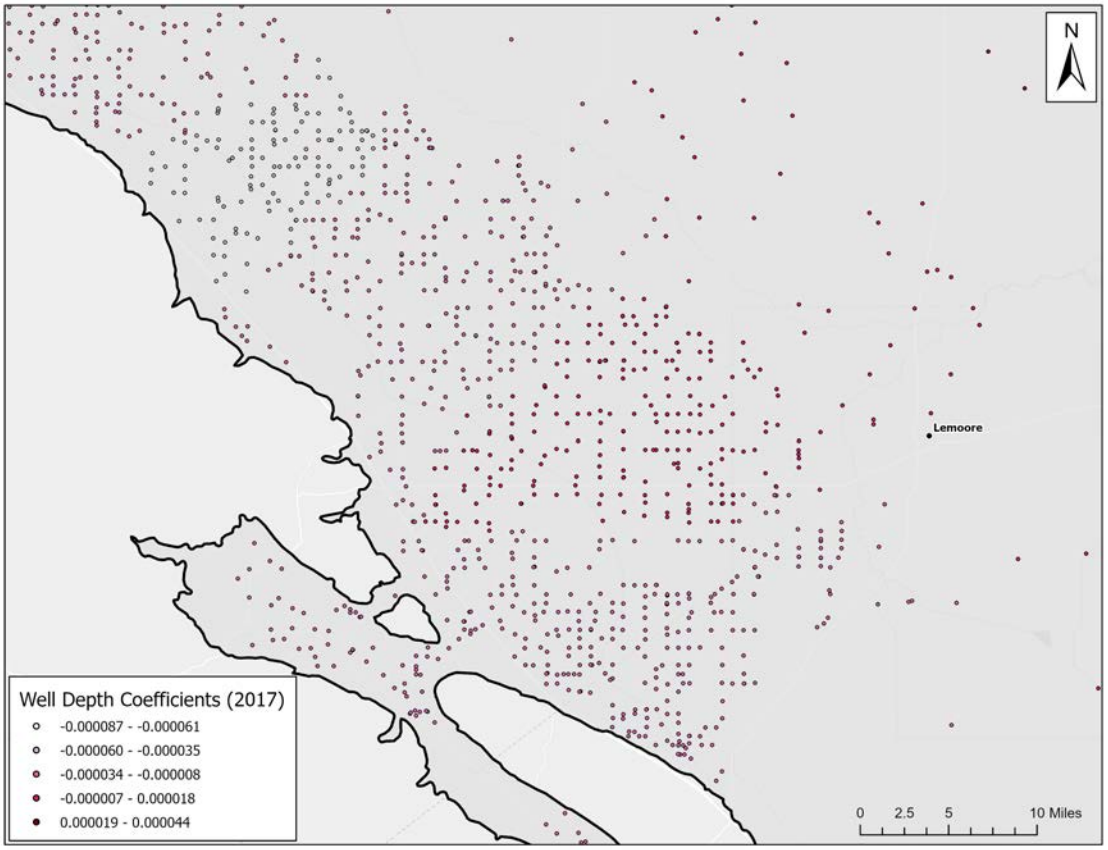


Figure 23. Distribution of well depth coefficients on the westside

It could go without saying, but this relationship is likely attributed to the water industry chasing the groundwater levels deeper and deeper. Without any amount of recharge occurring in the aquifer, largely due to drought, subsidence will continue to increase as lower groundwater levels are chased.

4.4.2. Geographically Weighted Regression Coefficients of Determination

Mapped local R^2 values for each GWR model show that the largest variance in coefficient of determination appears along the far westside of the San Joaquin Valley. This would be in areas around and near Bakersfield and even further west on the Kettleman Plain

Subbasin. This is where the project AOI is not filled in for the project area due to Kettleman Hills and Kettleman Dome.

For the 2015 model, the lowest R^2 values are located to the east of Fresno and to the immediate west of Bakersfield. There are several locations on the far west of the valley, such as in the Kettleman Plain Subbasin, where low R^2 values may also be found as shown in Figure 24.

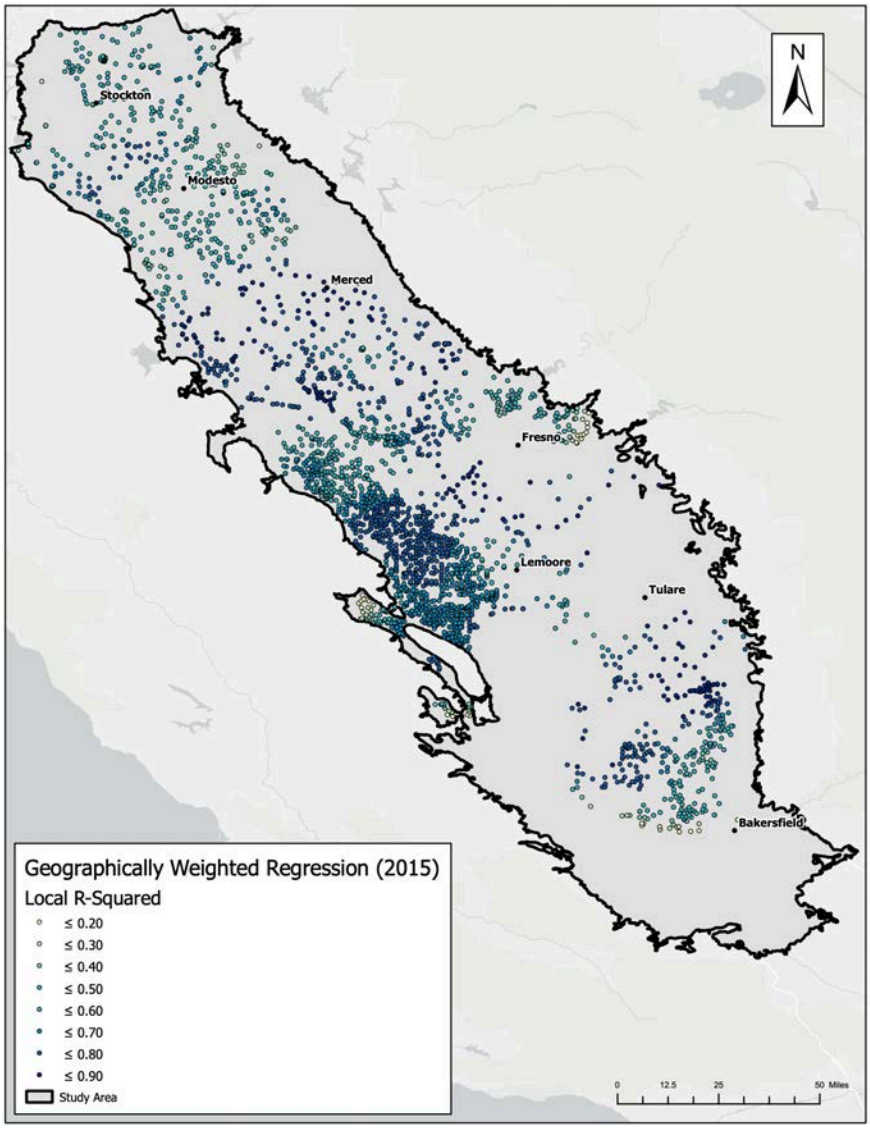


Figure 24. Local R-squared GWR model value maps, 2015

For the 2016 model, the highest R^2 values are located northwest of Lemoore and south of Merced and are exhibited in Figure 25. Such high values appear to be highly clustered and have the tendency to be > 0.70. The lowest R^2 values of the 2016 GWR model are located east of Fresno and on the far west of the study area in the Kettleman Plain Subbasin. South and west of Bakersfield has a tendency for R^2 values to be around 0.50. It may be noted that this area was also an area where 2016 SAR subsidence values showed little variance and few related values.

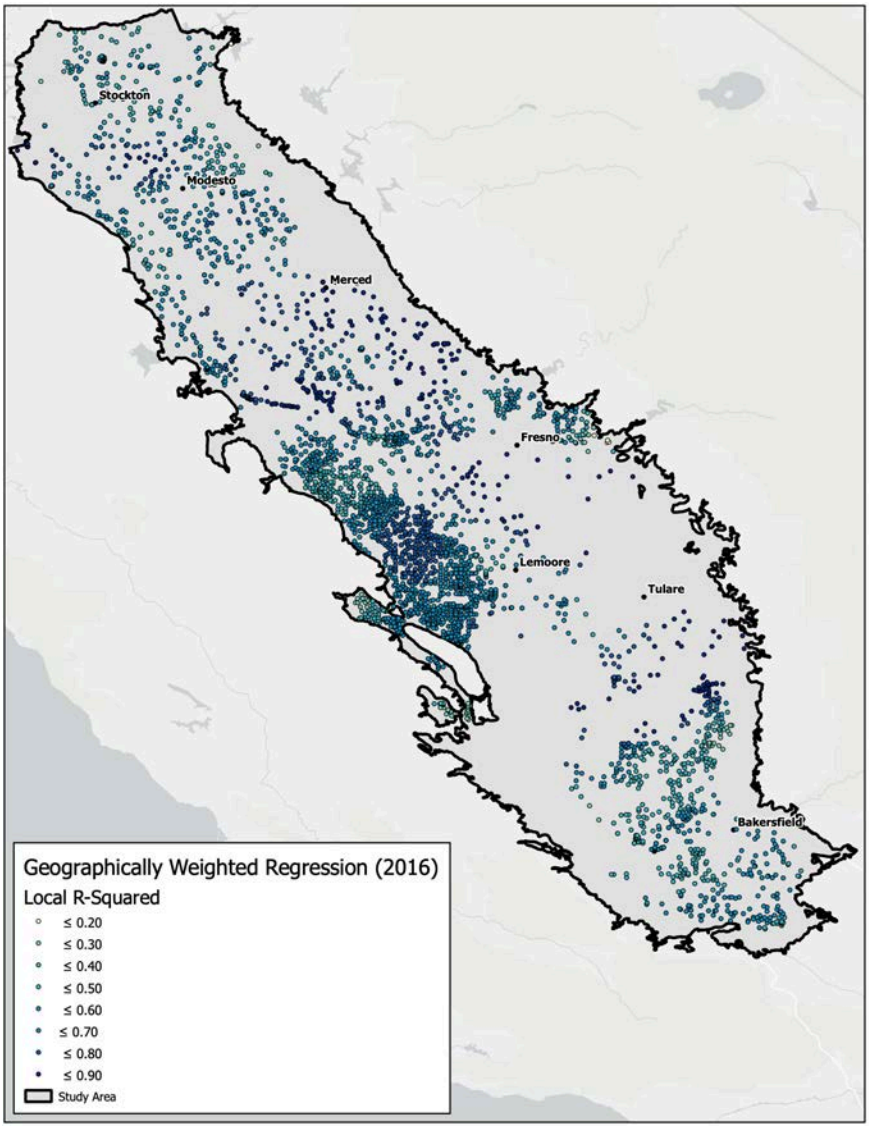


Figure 25. Local R-squared GWR model value maps, 2016

For the 2017 model, the highest R^2 values may be found throughout the central portion of the San Joaquin Valley (i.e. between Fresno and Lemoore up into Modesto). These higher values tend to be >0.80. One may note that there are higher values (>0.80) follow a similar trend southwest of Tulare as shown in Figure 26.

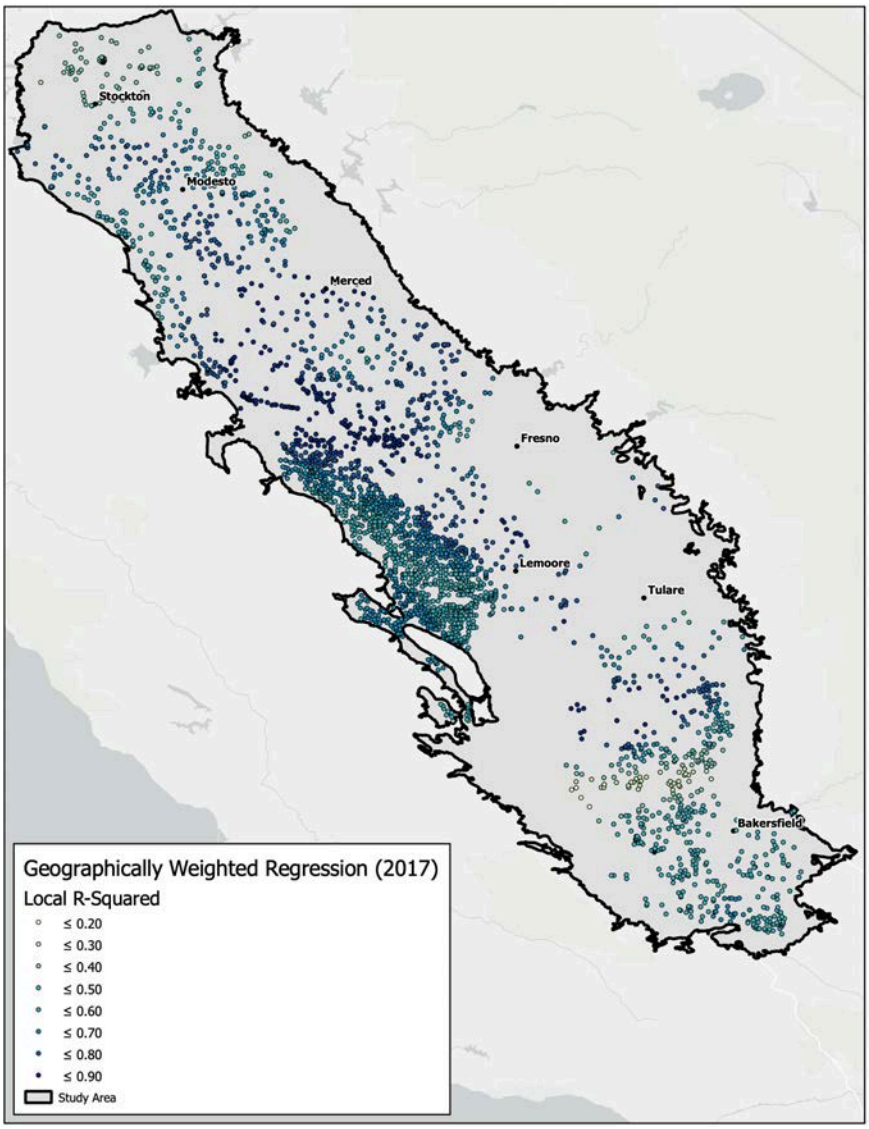


Figure 26. Local R-squared GWR model value maps, 2017

The lowest values of the 2017 model are located northwest of Bakersfield and are <0.30. Some lower values (<0.60 >0.50) may be found west of Lemoore and show some spatial

variability despite having had large, high value clusters in the previous two years' GWR results. This also complements the Anselin Local Moran's I cluster and outlier analysis from the previous section of this chapter. Figure 27 shows the statistical distribution of 2017 values.

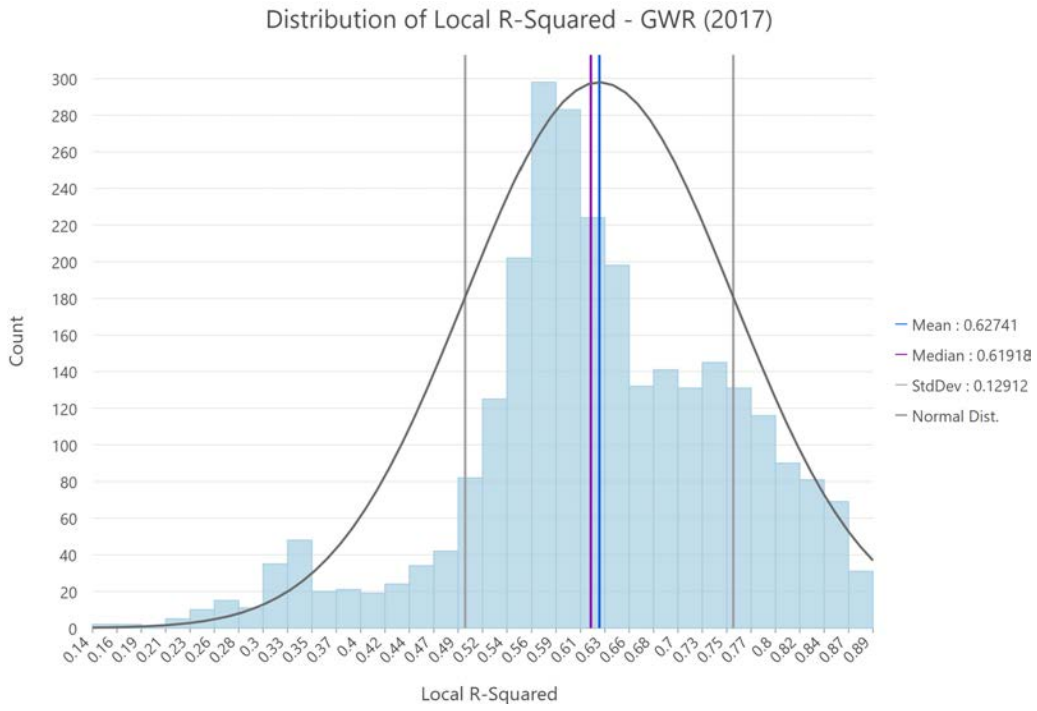


Figure 27. Histogram of GWR local R-squared values, 2017

For the 2018 GWR model, the highest R^2 values are located, once again, west of Lemoore. For the 2018 dataset, GWR R^2 values are higher (>0.80) on the interior of the clustered datapoints near Lemoore. The highest R^2 values (>0.90) appear down the central axis of the valley. Figure 28 shows the spatial distribution of the 2018 coefficients.

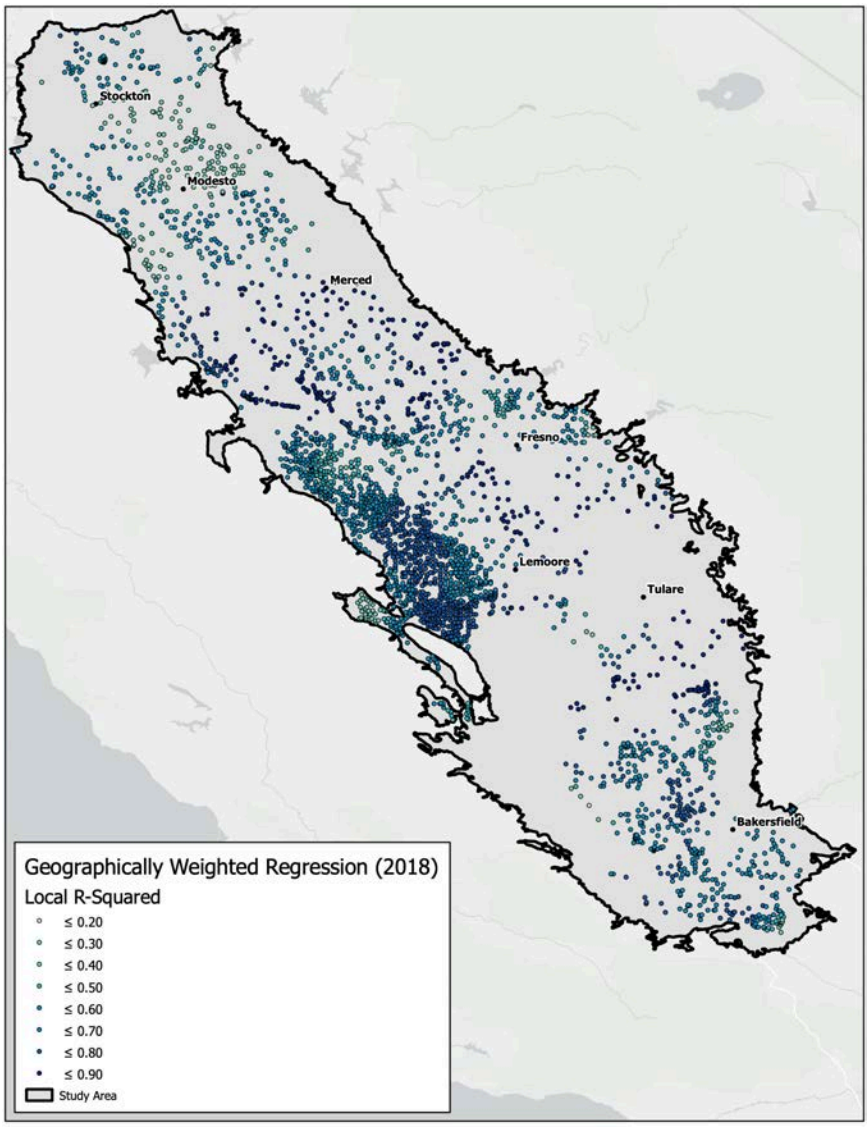


Figure 28. Local R-squared GWR model value maps, 2018

The lowest values (<0.30) of the 2018 model are located between Stockton and Modesto. The Bakersfield area is not exempt from local low values as that region yielded values that are similar to the 2017 GWR model. These values tend to be <0.60 on the southside of Bakersfield and <0.30 to the north.

For the 2019 model, there were fewer data points utilized in the analysis, thus yielding a larger distribution of the coefficient of determination, as shown in Figure 29. The highest R^2

values are located around Merced. Some higher values (>0.80) may be found north of Lemoore, but values that are <0.50 are also closely tied into the area making for a larger distribution of variability among R^2 values for 2019.

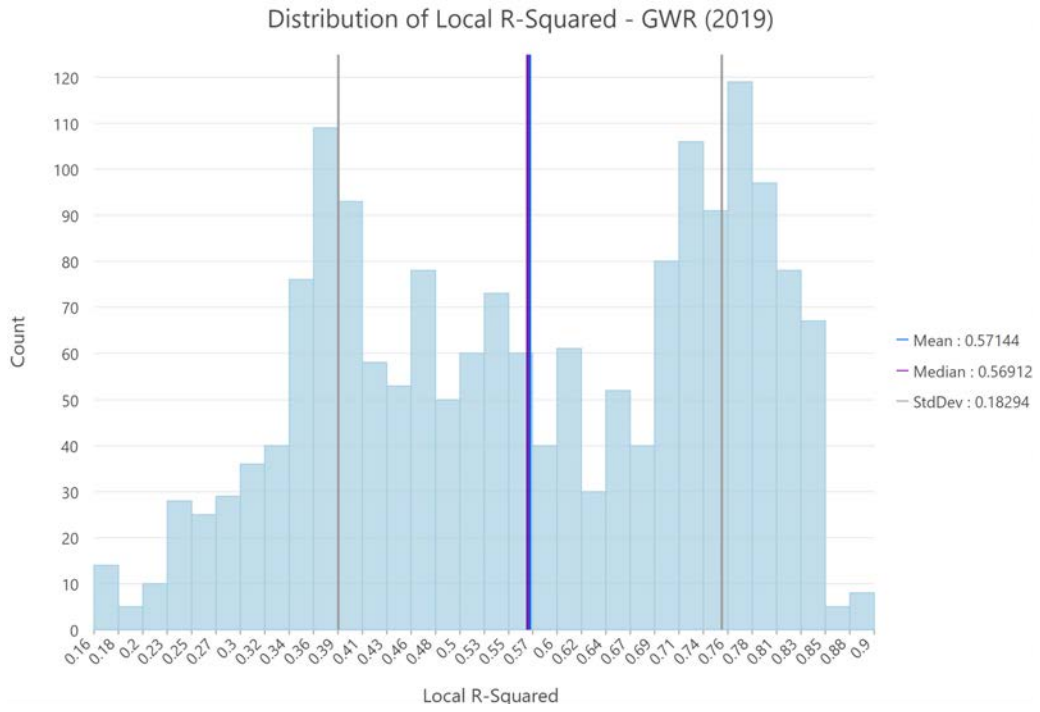


Figure 29. Histogram of GWR local R-squared values, 2019

The lowest values of the 2019 model are located around the Bakersfield area. Many of these values echo what was previously established in prior years (e.g. 2016, 2017, and 2018). A new area of low values (<0.20) shows up to the southwest of Modesto in the 2019 GWR model as shown in Figure 30.

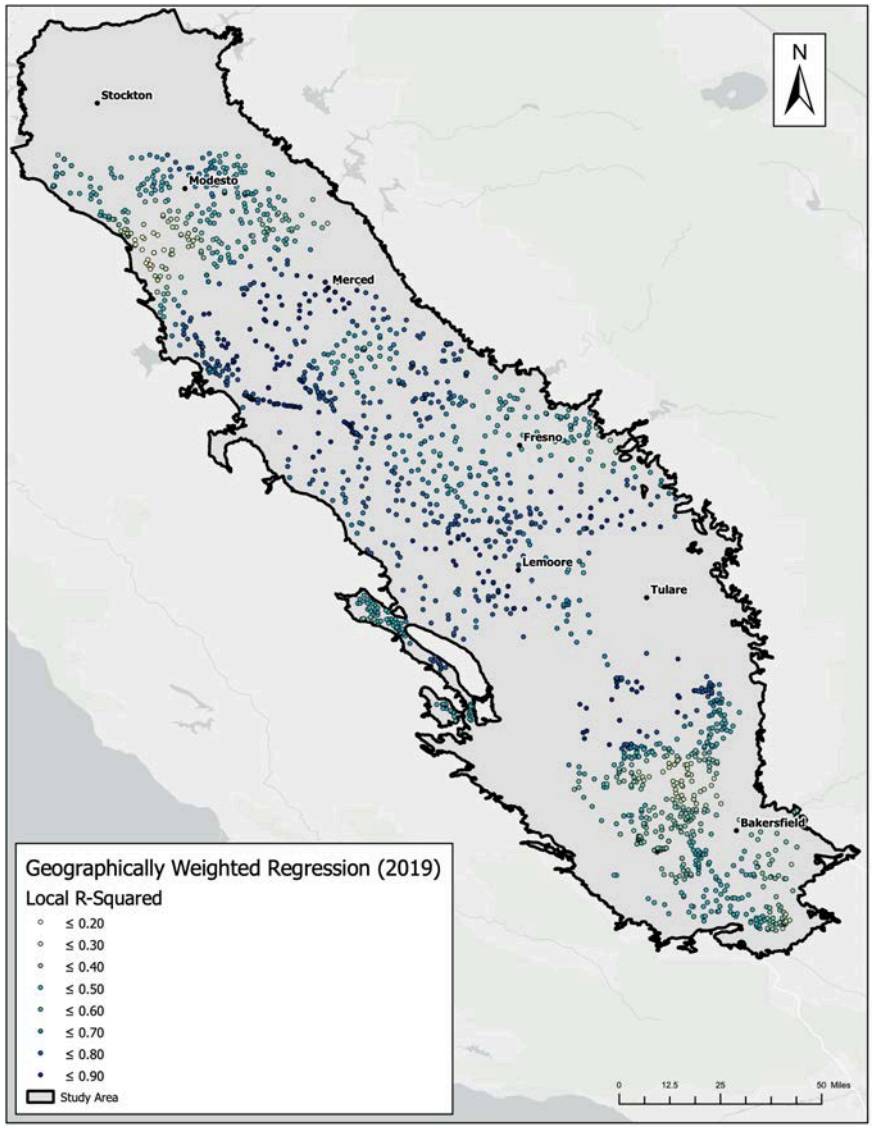


Figure 30. Local R-squared GWR model value maps, 2019

For the 2020 model, the highest R^2 values (>0.80) are located between Fresno and Lemoore as Figure 31 demonstrates. There are some relatively higher values north of Bakersfield and south of Tulare. One thing to note here is that there are next to no data around Tulare to the east and to the west. Once again, fewer data points were collected likely due to the onset of the COVID-19 pandemic. Groundwater levels were not collected during that time, or at least wells

have no records, due to the pandemic. The lowest R^2 values of the 2020 model are again found south of Bakersfield.

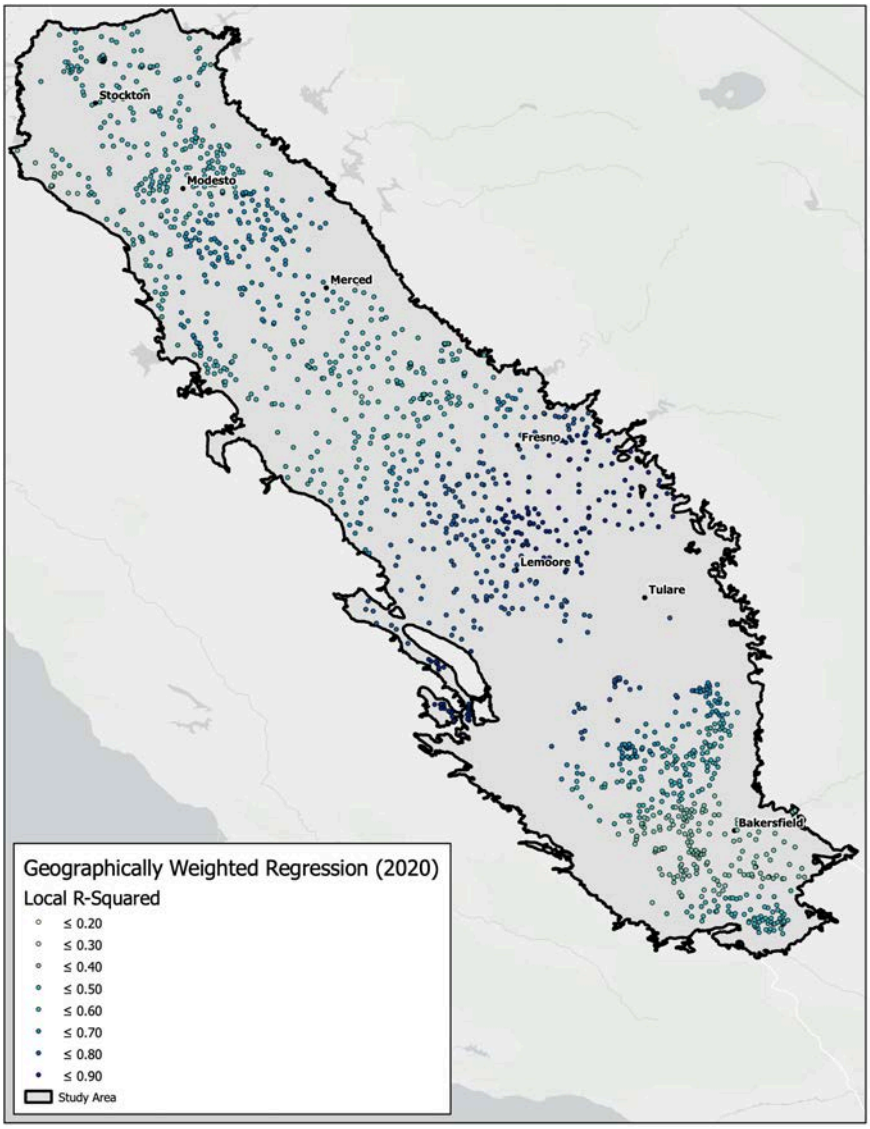


Figure 31. Local R-squared GWR model value maps, 2020

For the 2021 model, the R^2 values are very similar to the 2020 dataset and output as shown in Figure 32. The highest R^2 values (>0.80) are again found between Fresno and Lemoore. The similar missing data points should be noted as with the prior year, although visually, there several more points around Tulare in 2021. The only other notable change appears around

Bakersfield; in particular, to the south, where values have improved to >0.60. The lowest 2021 values (<0.3) have improved since 2020 from <0.20 to <0.30. These values may be found between Stockton and Modesto.

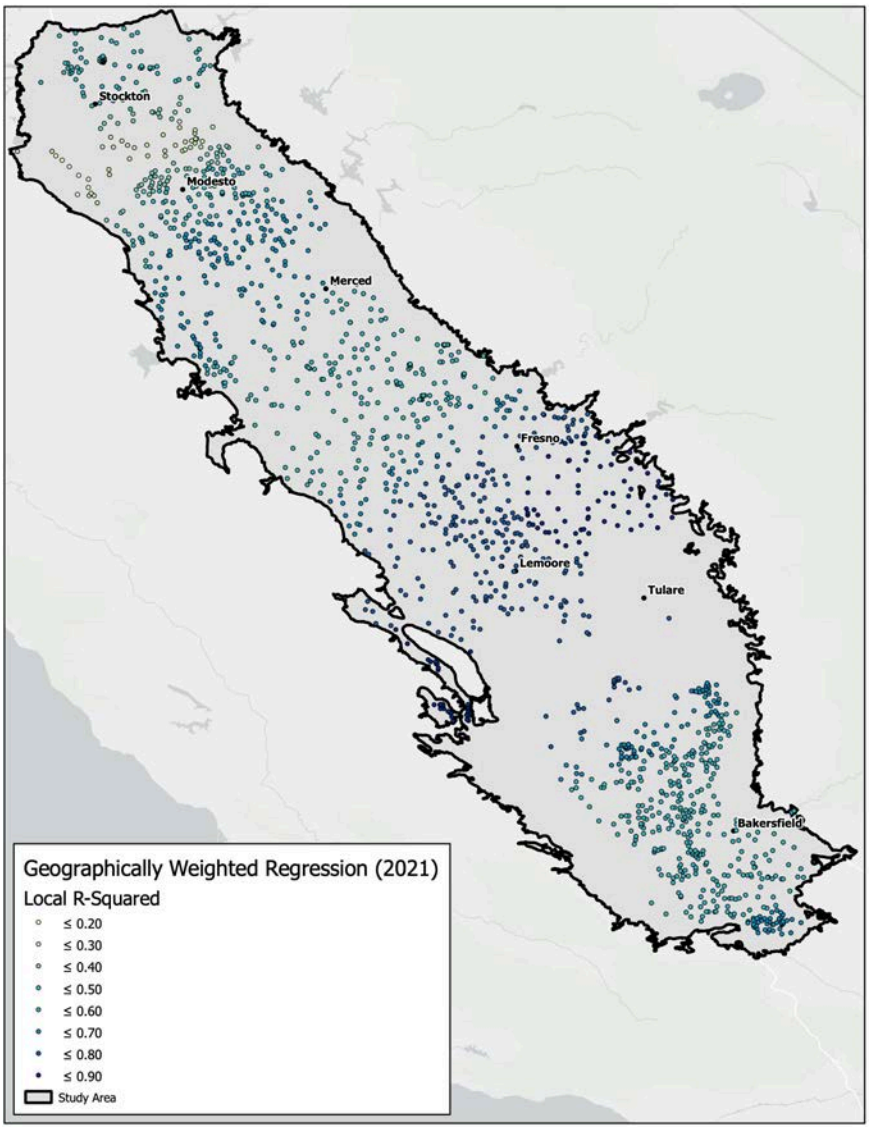


Figure 32. Local R-squared GWR model value maps, 2021

Model accuracy, based on coefficients of determination, was poorest for all years in the southern portion of the San Joaquin Valley, near Bakersfield. Since 2018, model accuracy was found to be poor along the on the northwest edge of valley just west of Modesto. The final area

of low model accuracy may be found near Kettlemen City. GWR model accuracy was highest (local R2 >.90) west of Lemoore as well as in the central region of the valley between Lemoore and Fresno. The central region of the valley averaged 92%, well above that of OLS at 28% for all years assessed. This area coincides with higher rates of subsidence, as well as lower groundwater level values. The existing variation among independent variables may account for the accuracy in the central region and even west of Lemoore. This is noted as locally weighted regression requires a certain amount of spatial variation in the independent variables, which may also explain the pattern of poor performance in areas like that of Modesto and Bakersfield.

This is complemented by the mapped standardized residuals shown in Figure 33. Note how the standardized residuals with higher standard deviations (red and dark blue) generally indicate that key variables are missing. However, as these data were engineered to have all strong performing variables, it is more likely that sudden, unexpected spatial variation has appeared among locations with higher standard deviations. This has likely led to the model incorrectly predicting the output (Esri 2022).

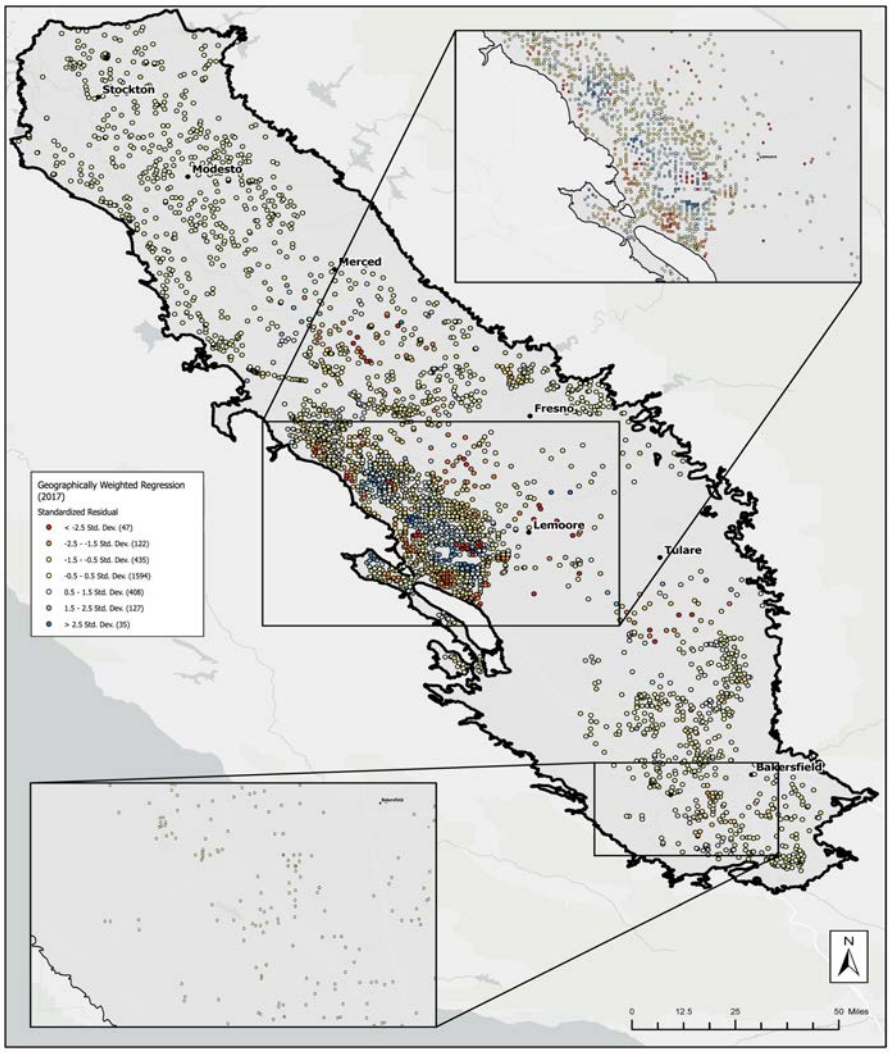


Figure 33. GWR standardized residuals, 2017

4.5. Model Performance and Assessment

As a measure of model performance, AIC_C values were assessed and recorded. Both MLR and GWR models for each year assessed yielded AIC_C. These, along with coefficients of determination, are the key assessors used in this study to gauge each model's accuracy.

4.5.1. AIC_C Assessment and Performance

The AIC_C values are a way to determine how well a model fits the data the model came from (Bevans 2022). As Table 12 shows, GWR outperformed MLR by almost triple the AIC_C values year-to-year.

Table 12. AIC_C performance by model and year

Year	MLR AICc	GWR AICc
2015	-1616.50466	-6765.8909
2016	-2337.789855	-7128.8269
2017	-3802.428091	-6913.942
2018	-4387.837165	-10594.4572
2019	-4455.49805	-7120.5861
2020	-2437.500709	-4096.7792
2021	-1894.767591	-3385.6421

While both MLR and GWR have had the same number of variables, and the exact same variables, it should be noted that each has been well grounded with hydrogeologic, engineering, and statistical knowledge and best practices.

4.5.2. Coefficient of Determination Assessment and Performance

Up to this point R^2 values have been discussed on a one-off basis. In fact, the focus on the example dataset for 2017 has been to compare both OLS and GWR models. Such a comparison between these two models has shown a drastic increase in goodness-of-fit. To echo this once more, the multiple R^2 value for the 2017 MLR model was 0.43417 (adjusted R-squared of 0.428766). The R^2 value of the 2017 GWR model was 0.8384 (adjusted R-squared of 0.8227). The improvement between the models is almost two times what it was initially. Table 13 further emphasizes this for each annual dataset and for each model.

Table 13. Comparison of coefficient of determination for MLR and GWR models

Year	MLR R^2	MLR AdjR^2	GWR R^2	GWR AdjR^2
2015	0.234182	0.232276	0.9044	0.8871
2016	0.25679	0.255203	0.8956	0.8773
2017	0.430417	0.428766	0.8384	0.8227
2018	0.157312	0.155129	0.9106	0.8954
2019	0.078938	0.074756	0.8439	0.8154
2020	0.366101	0.36316	0.8033	0.7917
2021	0.432287	0.429713	0.7987	0.7869

The average R^2 value for MLR is 0.27943243 (adjusted R^2 of 0.27700043). The average GWR R^2 value is 0.85641429 (adjusted R^2 of 0.8395). One might note that MLR had a particularly difficult time fitting the data to a global linear regression model (R^2 of 0.078938). Whereas the GWR model for 2019 had the fourth highest local R^2 value (0.8439).

4.5.3. Visual Comparison of Model Results

This next section goes back to the SAR data provided by the DWR and takes a look at the actual and predicted values in raster form. As previously outline, this approach is done knowing that there are other layers of complexity and error that may be introduced by adding interpolation of the predicted values to the outputs of this study. As also outlined in the methodology, values that were deemed to be "good" or "sufficient" (i.e. R^2 >0.60) were used to generated interpolation products. Figure 34 displays the IDW interpolation of land subsidence from 2017 GWR predicted values.

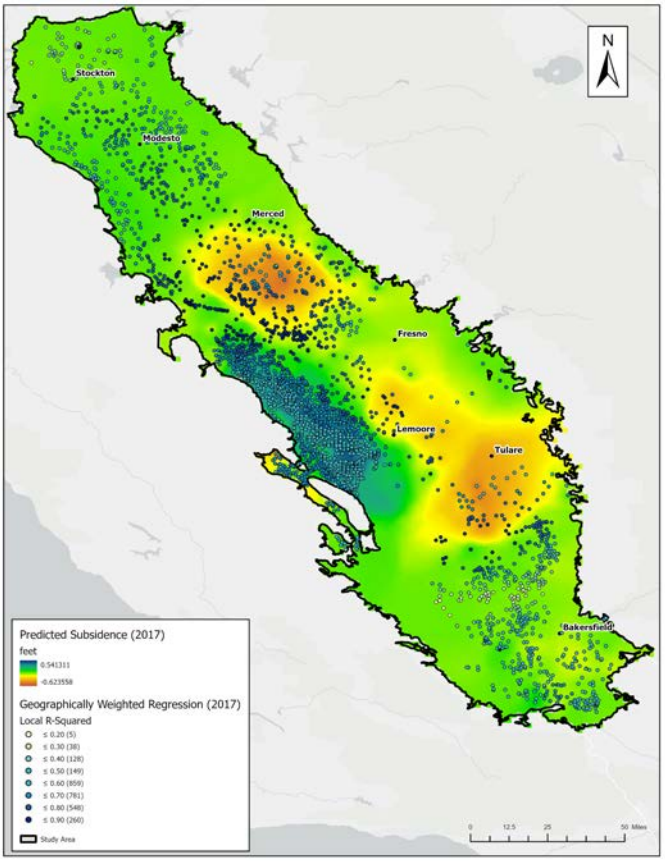


Figure 34. IDW interpolation from GWR predictions, 2017

Figure 34 also displays the overall count for each local R^2 value in the legend. That count may be found in parentheses. For the 2017 dataset, only 5 points (<1%) had an R^2 value < 0.20. Notably, 88% of the total count of R^2 values are >0.60. Based on the accompany R^2 values, there appears to be good data coverage throughout the San Joaquin Valley. Enough that perhaps, one might feel confident in the small number of lower R^2 values that are scattered throughout the valley.

In Figure 35, the predicted and interpolated output is put next to the original 2017 SAR data that the dependent variable was initially derived from. The left map in Figure 35 is from the 2017 GWR model of which the model's predicted values have been interpolated from the IDW

algorithm. The map on the right is of the original 2017 SAR raster dataset of which the land subsidence (dependent variable) was derived from. All SAR maps of land subsidence may be found in Appendix E – Original SAR Land Subsidence Maps. Interpolated GWR predicted land subsidence maps for years 2015-2021 may be found in Appendix J – GWR Predicted Land Subsidence Maps.

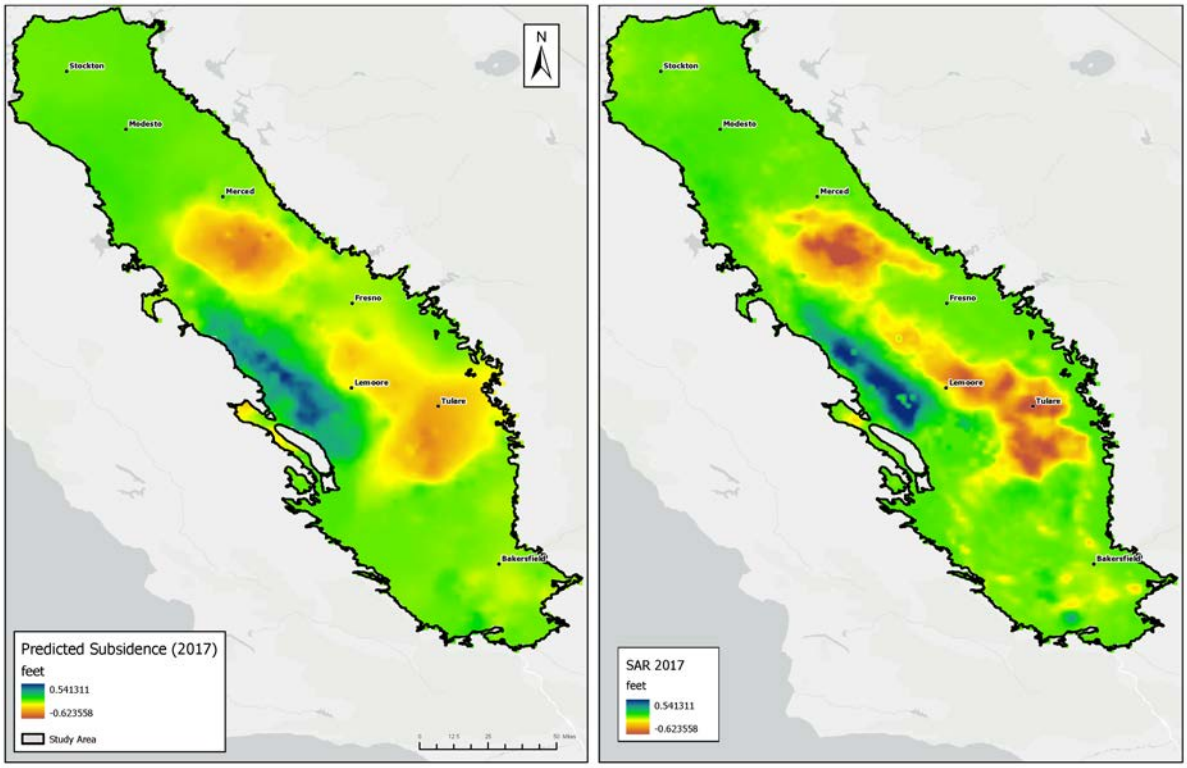


Figure 35. 2017 GWR and SAR land subsidence maps

Qualitative observations about these two maps would lead one to note that there appears to be an overprediction of land subsidence around the Fresno area in 2017. While the R^2 values in the area are >0.80, many are in fact >0.90, yet the brighter reds and yellows signifying uplift bleed into the east of the study area. This is a great example of a secondary error from interpolating predicted values to great a raster. Nonetheless, Figure 36 shows the quantitative difference between these two rasters.

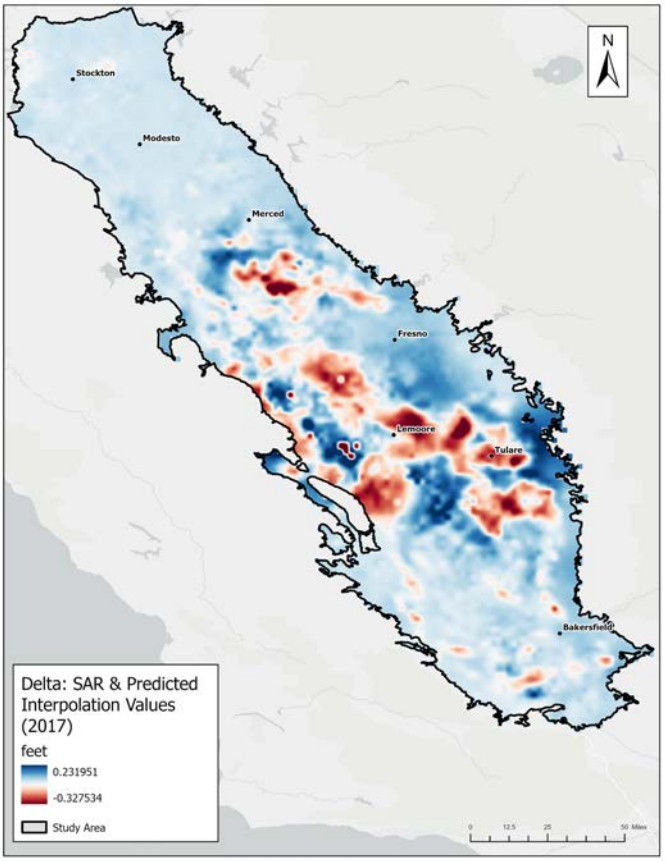


Figure 36. Difference map of SAR and GWR predictions, 2017

The absolute maximum difference between the 2017 SAR data and the interpolated GWR prediction values is |0.33 feet|. This value is a negative value that would show a maximum difference of -0.33 feet by which the GWR prediction and interpolation is overestimating the amount of subsidence. The average difference between these rasters is -0.0019 feet. Figure 37 displays the distribution of the delta values between these two rasters. One may note a very normal distribution in delta values. Additionally, two standard deviations contain most of the delta values. Such values are between -0.05 feet and 0.06 feet.

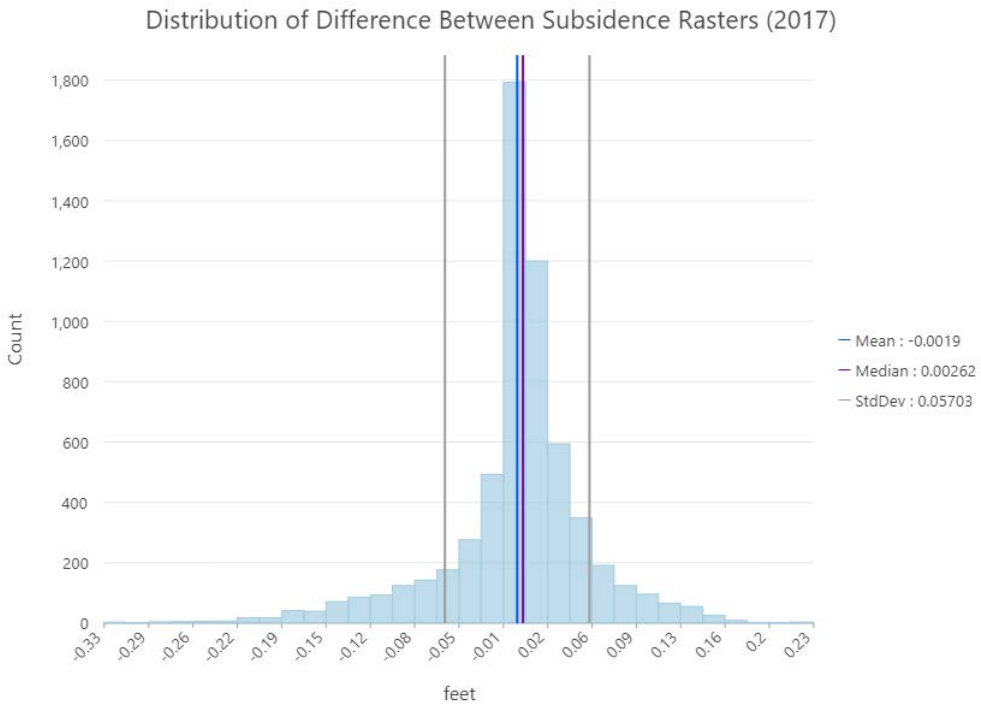


Figure 37. Histogram of delta values between subsidence rasters, 2017

Once again, it must be noted that this approach adds two levels of error to the predictions and requires further investigation for the best practice in interpolating predicted values from a spatial regression model. Such was not in the scope of this study.

Chapter 5 Discussion and Conclusions

Patterns and spatial distributions of land subsidence in the San Joaquin Valley were examined and explored through multi-variate analyses, global regression models, and local regression techniques. The use of experiential knowledge pertinent to hydrogeology and engineering practices were combined with spatial regression modeling techniques to yield an effective, easy-to-update, and accurate model for assessing land subsidence patterns in California's San Joaquin Valley.

GIS spatial analytics and exploratory data analyses of land subsidence in relation to key hyrdrogeologic and engineering variables such as total well depth, statistical distribution of fine-grained sediment, and the presence of a large aquitard such as the Corcoran Clay all have environmental impacts on predicting the spatial patterns of land subsidence. While each of the nine variables assessed in this study prove to be influential, global regression methods give insight to spatial autocorrelation and spatial clustering of similar values as well as the identification of outliers. Quantitative and qualitative analyses of the spatial distribution of variable coefficients have identified regions in the San Joaquin Valley that may be light in data quantity and quality. Such regions have also been key in identifying where local regression models may fall short and where global regression models completely missing the mark when it comes to predicant spatial patterns in land subsidence. Based on the requirements of the Sustainable Groundwater Management Act (SGMA), local spatial regression models can assist Groundwater Sustainability Agency (GSAs) and Groundwater Sustainability Plans (GSPs) with the management and use of groundwater to avoid the potential "undesirable result" of land subsidence.

This chapter summarizes the results, shortcomings, and potential future solutions to finetune spatial regression models for predicting land subsidence. Regression model results and areas of improvement are discussed and related to areas identified for improvement and/or model adaptations.

5.1. Hydrogeologic and Engineering Impacts

Previous studies have not fully integrated experiences from the practice of hydrogeology and engineering with that of academic concepts. The methodology in this study have integrated the best of both worlds with what may be considered a successful and reliable outcome. As demonstrated with the AIC_C values from both global and local spatial regression models, including each identified variable is key in grounding the statistical concepts with real world application. One may note that the inclusion of each independent variable yielded a higher AIC_C for the spatially lagged model as well as the GWR model. AIC_C had little to no improvement as the suggested independent variables were removed. In fact, the best performing AIC proliferated from the 2019 GWR model with an AIC of -8208.716. With poorest performing AIC stemming from the 2015 OLS model with an AIC of -1616.5047, it goes to show the large difference the models make, but also what difference the key independent variables play.

When assessed at the local scale, subsidence is based on a multitude of local factors, including overexploitation, water level drawdown, geology, and water year type (California Department of Water Resources 2022). These sedimentary deposits are comprised of unconsolidated gravel, sand, silt, and clay that define the Tulare Formation (Hill 1964). Additionally, numerous lenses of fine-grained sediments (e.g. silt, sandy silt, sandy clay, and clay) are also present and according to Page (1973) make up over 50% of the total aquifer thickness.

Most of the fine-grained materials have been mapped using geophysical logs, seismic surveys, and drill core throughout the San Joaquin Valley. Most notably, and most spatial widespread, is the lithologic unit of the Corcoran Clay Member of the Tulare Formation that exists along the majority of the westside of the valley (Bertoldi et al. 1991). The Corcoran Clay Member is a key component of groundwater hydraulics in the valley. Lees et al. (2021) have subdivided the Tulare Formation, the primary aquifer throughout the valley, into three different hydrostratigraphic layers: the unconfined to semi-confined upper Tulare (or upper aquifer), the Corcoran Clay Member, and the lower Tulare (or confined lower aquifer). Associating groundwater extraction and the ramifications of land subsidence within the Tulare Formation would be impossible without identification of confined and unconfined portions of the aquifer system.

Ali et al. (2020) and Chu et al. (2021) had even gone so far as to differentiate among the confining layers in their study. Having this understanding and knowing that aquitards have such a large impact on groundwater systems, it makes sense to include such layers and their spatial distribution throughout the San Joaquin Valley in any land subsidence analysis.

Furthermore, including the Corcoran Clay thickness and percent of fine-grained material within the Tulare aquifer itself lends itself to the engineering of differential compaction that was mentioned in the introduction of this study. Including the general principle of compaction as it occurs when sediment is unconsolidated and has high percentages of fine-grained sands, silts, and clays cannot and should not be removed from the equation when assess land subsidence (Davis and Poland 1957; Davis et al. 1964; Poland et al. 1975). Yes, even the spatial distribution of fine-grained material and thickness of aquitards have lent to a better understanding of the spatial patterns of land subsidence within the San Joaquin Valley.

5.2. Regression Successes

Once the key variables that are geologically and engineering based were established, the exploration of regression models was able to be established. The process of establishing a regression model was accomplished through relations among variables per Chu et al. (2021) and Ali et al. (2020). As noted in this study's methodology, a key difference between this study and those emulated includes the use of more variables that are based both in the real world and in the practice of groundwater extraction. The relationship between annual subsidence the independent variables was formulated, and the spatial patterns of subsidence were subsequently estimated. With this, local regression in the form of GWR performed better than global regression in the form of MLR. This was anticipated based on the work of Chu et al. (2018). The R^2 and AIC values demonstrated that locally weighted regression performs in a manner that is sufficient for estimating patterns of land subsidence throughout the San Joaquin Valley.

Low coefficients of determination continually appeared in the GWR model on the far west boundary of the study area, as well as to the south end of the valley to the west and to the south of Bakersfield. In the southern portion of the valley, one may note that annual SAR land subsidence values tend to be very similar, almost to the point that the grid values for each year's raster are not correctly represented. This apparent lack in data quality may be what led to low coefficients of determination in this region.

The local regression model (GWR) coefficient maps demonstrated large and consistent patterns of land subsidence through the central axis of the valley. This same segway tends to hold the higher R^2 values for each annual local regression model. It should also be noted that inflation to the west of Lemoore also shows strongly in large clusters. These same clusters tend to exhibit R^2 values that are > 0.70.

The results of the local and global regression models, when compared, are similar to that of Chu et al. (2021), Chu et al. (2018) and Ali et al. (2020). While these authors utilized the relationship between land subsidence and groundwater level change through time, groundwater level change demonstrated multicollinearity and was not fully assessed at the local level in this study. Furthermore, due to the direct, linear connection between the percent of coarse-grained sediment and the percent of fine-grained sediment, coarse-grained sediment was removed from the local regression analysis. This variable also demonstrated multicollinearity when constructing a GWR model in ArcGIS Pro. Other variables that had similar challenges were the top perforation depth and base perforation depth. Both variables were directly connected to each well's completion length. As mentioned in this study's methodology, the completion length variable was calculated from top and base perforation depths. However, each of these three variables did not show multicollinearity when looking at VIF results. Each VIF value was larger than 1.0. ArcGIS Pro was able to catch the redundancy of each removed variable that would have otherwise led to an unstable regression model (Esri 2022).

While Chu et al. (2021) utilized IDW to generate a raster of their predicted land subsidence values, a secondary level of error is being introduced through the use of such interpolation methods. As previously mentioned, the purpose of this study is not to explore concepts of interpolation as it pertains to groundwater modeling or land subsidence patterns. Rather, the use of IDW in this study follows a consist method by which predicted land subsidence values have been explored (Chu et al. 2018; Chu at el. 2021; Ali et al. 2020). Interpolation methods allow a qualitative look at model results and can be used to visually assess strong and weak performance in a spatial regression model. This is especially true when compared side-by-side with the SAR land subsidence dataset.

Taking observations around interpolated surfaces further, the generated delta rasters of this study also show areas where either more data are needed or where the underlying model data need to be refined. Such refining may come in the form of a larger clustering of groundwater measurements in specific areas (e.g. west of Modesto). Worded in another way, more spatial clustering of groundwater level measurements may help establish better predictions with the spatial regression model. This would show stronger results in both the R^2 values and in interpolation methods in those areas that currently demonstrate relatively weak performance.

The exploration of homoscedasticity and the ability for GWR to assess dependent variable variation allows one to contrast seemingly small changes from one groundwater well to the next while attaining high confidence in predicting land subsidence. This study's approach can be used to show spatial patterns of land subsidence through space and time even as each independent variable changes through that same space and time.

5.3. Further Development

The results of spatial analysis and local spatial regression models demonstrated that even on a well-to-well basis there are large differences in land subsidence patterns. For this reason, it is strongly suggested that the nine subbasins be assessed separately from the San Joaquin Valley as a whole. This became very clear when low R^2 were showing up south of Bakersfield as well as on the western edge of the valley near Kettlmen Hills. Make such subdivisions not only caters to local GSPs but also may enable modelers to incorporate localized aquitards that are not as spatially widespread as the Corcoran Clay. This also implies that more localized hydrogeological systems and variables, if the data are readily available, may be incorporated into the spatial regression model and yield higher confidence in the coefficient of determination for each well location.

In the spirt of continuous improvement that this study was based on, there are several key items that might be included in future analyses. One such variable would be the inclusion of volumes of water extracted at groundwater well locations. It is unfortunate that such data are not readily available via DWR, RWQCB, USGS, or GAMA. As subbasin through the San Joaquin Valley work on meeting the statutes of SGMA, such volumes are likely to become easier to access and may be an even more impactful independent variable in predicting land subsidence.

Furthermore, working with groundwater agencies, such as DWR and RWQCB, to assess recharge effects and hydrodynamic lag based on groundwater draw down (i.e. groundwater level change through time) and resulting subsidence could be considered. This comes at a risk of overcomplicating even local regression models. Thus, the effectiveness of using few, but well established, independent variables should still relevant. Doing so will guarantee an easy to update, but highly accurate model for predicting patterns of land subsidence.

5.4. Conclusion

The goal of this study was to explore and predict the spatial distribution of land subsidence through spatial regression models in California's San Joaquin Valley. Spatial data analysis incorporated autocorrelation that is masked by typical statistical analyses. Additionally, a global regression model failed to incorporate variability among independent variables throughout the valley. Spatial analyses at a local scale provides insight to spatial variability among variable coefficients.

The use of a spatial regression model within a GIS enables GSAs to incorporate better land subsidence predictions into their groundwater sustainability plans. Incorporating newly drilled wells, their associated depths, completion lengths, as well as notation of completion in the upper or lower Tulare makes spatial regression model predictions easy to update and accurate

enough to make GSP adjustments with the valley and throughout subbasins. Further improvements that include refined completion intervals, updated groundwater measurements (as per SGMA), and the integration of volumes of water extracted may help GSAs avoid undesirable results. Therefore, this method can be used efficiently for land subsidence management and might be widely used for subsidence estimation solely based on experiential hydrogeology and engineering variables. As noted above, the quality of each independent variable affects the estimation accuracy of land subsidence. Yet this spatial regression model can be relevant in terms of SGMA and land subsidence regulation.

References

2021 California Code Water Code – WAT, DIVISION 6 - CONSERVATION, DEVELOPMENT, AND UTILIZATION OF STATE WATER RESOURCES. PART 2.74 - Sustainable Groundwater Management CHAPTER 1 - General Provisions Section 10720.

Ahmed, A.W, Kalkan, E., Guzy, A., Alacali, M., and Malinowska, A. 2020. "Modeling of Land Subsidence Caused by Groundwater Withdrawal in Konya Closed Basin, Turkey," Proceedings of the International Association of Hydrological Sciences 382, 382 (April). Gottingen: Copernicus GmbH: 397–401. doi:10.5194/piahs-382-397-2020.

Ali, Muhammad Zeeshan, Hone-Jay Chu, and Thomas J Burbey. 2020. "Mapping and Predicting Subsidence from Spatio-Temporal Regression Models of Groundwater-Drawdown and Subsidence Observations," Hydrogeology journal 28, 28 (8). Berlin/Heidelberg: Springer Berlin Heidelberg: 2865–76. doi:10.1007/s10040-020-02211-0.

Azarakhsh, Zeinab, Mohsen Azadbakht, and Aliakbar Matkan. 2022. "Estimation, Modeling, and Prediction of Land Subsidence Using Sentinel-1 Time Series in Tehran-Shahriar Plain: A Machine Learning-Based Investigation," Remote sensing applications 25, 25 (January). Elsevier B.V. doi:10.1016/j.rsase.2021.100691.

Baguley, T. 2012. "Serious stats: A guide to advanced statistics for the behavioral sciences." Palgrave Macmillan, 2012. P. 402.

Barnett, B., Townley, L.R., Post, V, Evans, R.E., Hunt, R.J., Peeters, L., Richardson, S., Werner, A.D., Knapton, A., and Boronkay, A. 2012. Australian groundwater modelling guidelines, National Water Commission, Canberra, June, 191 p. http://archive.nwc.gov.au/library/waterlines/82

Bawden, G.W., Sneed, Michelle, Stork, S.V., and Galloway, D.L., 2003, Measuring human-induced land subsidence from space: U.S. Geological Survey Fact Sheet 069–03, 4 p. http://water.usgs.gov/pubs/fs/fs06903/.

Belgiu, Mariana, and Lucian Drăguţ. 2016. "Random Forest in Remote Sensing: A Review of Applications and Future Directions," ISPRS journal of photogrammetry and remote sensing 114, 114 (April). Elsevier B.V: 24–31. doi:10.1016/j.isprsjprs.2016.01.011.

Bertoldi, G.L., Johnston, R.H., and Evenson, K.D. 1991. Ground Water in the Central Valley, California - A Summary Report: U.S. Geological Survey Professional Paper 1401-A.

Bevans, R. (2022, May 25). *Akaike Information Criterion | When & How to Use It (Example).* Scribbr. Retrieved November 3, 2022, from https://www.scribbr.com/statistics/akaike-information-criterion/

Bivand, R., Gomez-Rubio, V., and Pebesma, E. 2013. *Applied Spatial Data Analysis with r, Second Edition*. New York: Springer.

Bivand R, Millo, G., Piras, G. 2021. "A Review of Software for Spatial Econometrics in R." *Mathematics*, 9(11). doi:10.3390/math9111276. https://www.mdpi.com/2227-7390/9/11/1276

Bolstad, P. 2016. *GIS Fundementals: A First Text on Geographic Information Systems.* 6th ed. Action, MA: XanEdu. p. 392-394.

Breusch, T.S. and Pagan, A.R. 1979. "A Simple Test for Heteroscedasticity and Random Coefficient Variation." *Econometrica* 47, 1287–1294

Buis, A. and Thomas, T. "NASA Data Show California's San Joaquin Valley Still Sinking." *NASA*, Feb. 28, 2017. https://www.nasa.gov/feature/jpl/nasa-data-show-californias-san-joaquin-valley-still-sinking

Burbey, T.J. 2001. Stress-Strain Analyses for Aquifer-System Characterization. Groundwater, 39: 128-136. https://doi.org/10.1111/j.1745-6584.2001.tb00358.x

California Department of Food and Agriculture (CDFA). 2010. "California Agriculture Statistical Report 2008-2009."

California Depart of Water Resources. 2022. "New Data Shows Subsidence Continued in Water Year 2021, But Pace Slower than Seen in Previous Droughts." Accessed: May 24, 2022.

Castellazzi, P., Martel, R., Galloway, D.L., Longuevergne, L. and Rivera, A. 2016. Assessing Groundwater Depletion and Dynamics Using GRACE and InSAR: Potential and Limitations. Groundwater. 54. n/a-n/a. 10.1111/gwat.12453.

Chi, G. and Zhu, J. 2019. Spatial Regression Models for the Social Sciences. Thousand Oaks, CA: SAGE Publications.

Chu, H.J., Ali, M.Z., Tatas, and Burbey, T.J.. 2021. "Development of Spatially Varying Groundwater-Drawdown Functions for Land Subsidence Estimation," Journal of hydrology. Regional studies 35, 35 (June). Elsevier B.V: 100808. doi:10.1016/j.ejrh.2021.100808.

Chu, H.J., Kong, S.J., and Chang, C.H. 2018. "Spatio-Temporal Water Quality Mapping from Satellite Images Using Geographically and Temporally Weighted Regression," ITC journal 65, 65 (March). Elsevier B.V: 1–11. doi:10.1016/j.jag.2017.10.001.

Cromwell, J.B., Labys, W. C., and Terraza, M. 1994. *Univariate Tests for Time Series Models*, Sage, Thousand Oaks, CA, p. 20–22.

Davis, G.H., Lofgren, B.E., and Mack, S. 1964. Use of ground-water reservoirs for storage of surface water in the San Joaquin Valley, California: U.S. Geological Survey Water-Supply Paper 1618.

Davis, G.H. and Poland, J.F., 1957. Ground-water conditions in the Mendota-Huron area, Fresno and Kings Counties, California: U.S. Geological Survey Water-Supply Paper 1360-G, p. 409-588.

Esri. 2022. "Multicollinearity." Accessed November 9, 2022: https://pro.arcgis.com/en/pro-app/latest/tool-reference/spatial-statistics/h-multicollinearity.htm

Faunt, C.C. 2009. "Groundwater availability of the Central Valley Aquifer, California." USGS Professional Publication. Paper 1766.

Faunt, C.C., Sneed, M., Traum, J., and Brandt, J.T. 2015. "Water Availability and Land Subsidence in the Central Valley, California, USA," Hydrogeology Journal 24, 24 (3).

Fergason, K C, M L Rucker, and B B Panda. 2015. "Methods for Monitoring Land Subsidence and Earth Fissures in the Western USA." Proceedings of the International Association of Hydrological Sciences 372. Copernicus GmbH. doi:10.5194/piahs-372-361-2015.

Fetter, C. W. 1942-. 1994. *Applied Hydrogeology*. New York : Toronto : New York, Macmillan.

Fotheringham, A. S., Crespo, R., and Yao, J. (2015) Geographical and temporal weighted regression (GTWR). Geographical Analysis, 47(4), pp. 431-452. (doi:10.1111/gean.12071)

Galloway, D.L., Burbey, T.J., 2011. Review: regional land subsidence accompanying groundwater extraction. Hydrogeol. J. 19, 1459–1486.

Galloway, D.L, Erkens, G., Kuniansky, E.L., and Rowland, J.C. 2016. "Preface: Land Subsidence Processes," Hydrogeology journal 24, 24 (3). Berlin/Heidelberg: Springer Berlin Heidelberg: 547–50. doi:10.1007/s10040-016-1386-y.

Galloway, D.L. and Riley, F.S. 1999. San Joaquin Valley, California: Largest human alteration of the Earth's surface. U.S. Geological Survey Circular 1182. 1182. 23-34.

Garone, P. The Fall and Rise of the Wetlands of California's Great Central Valley. 1st ed. University of California Press, 2011. http://www.jstor.org/stable/10.1525/j.ctt1pp4f6.

Griffith, D. 1987. Spatial Autocorrelation: A Primer. Resource Publications in Geography, Association of American Geographers.

Guzy, A. and Malinowska, A.A. 2020. "State of the Art and Recent Advancements in the Modelling of Land Subsidence Induced by Groundwater Withdrawal," Water (Basel) 12, 12 (7). Basel: MDPI AG: 2051. doi:10.3390/w12072051.

Haining, R., Wise, S. and Ma, J. 1998. Exploratory Spatial Data Analysis. Journal of the Royal Statistical Society: Series D (The Statistician), 47: 457-469. https://doi.org/10.1111/1467-9884.00147

Hanak, E., Escriva-Bou, A.,Gray, B., Green, S., Harter, T., Jezdimirovic, J., Lund, J., Medellín-Azuara, J., Moyle, P., and Seavy, N. 2019. "Water and the Future of the San Joaquin Valley OVERVIEW."

Hanson, R.T., Flint, A.L., Faunt, C.C., Cayan, D.R., Flint, L.E., Leake, S.A., Schmid, W. 2010. "Integrated simulation of consumptive use and land subsidence in the Central Valley, California, for the past and for a future subject to urbanize and climate change." Conference Paper, California Water Science Center.

Harrell, F. E. 2015. Regression modeling strategies: With applications to linear models, logistic and ordinal regression, and survival analysis. Springer.

Hill, F. L. 1964. Harvester gas field: California Department of Conservation, Division of Oil and Gas, Summary of Operations, California Oil Fields, v. 50, no. 1, p. 11-15.

Hoffmann, J., Galloway, D.L., and Zebker, H.A. 2003. "Inverse Modeling of Interbed Storage Parameters Using Land Subsidence Observations, Antelope Valley, California," Water resources research 39, 39 (2). American Geophysical Union: 1031-n/a. doi:10.1029/2001WR001252.

Hubbard, S.S., and Rubin, Y. 2000. "Hydrogeological Parameter Estimation Using Geophysical Data: A Review of Selected Techniques." Journal of Contaminant Hydrology 45. Elsevier BV. doi:10.1016/s0169-7722(00)00117-0.

Jafari, F., Javadi, S., Golmohammadi, G., Karimi, N. and Mohammadi, K. 2016. "Numerical Simulation of Groundwater Flow and Aquifer-System Compaction Using Simulation and InSAR Technique: Saveh Basin, Iran," Environmental earth sciences 75, 75 (9). Berlin/Heidelberg: Springer Berlin Heidelberg: 1. doi:10.1007/s12665-016-5654-x.

Jeanne, P., Farr, T.G., Rutqvist, J., and Vasco, D.W. 2019. "Role of Agricultural Activity on Land Subsidence in the San Joaquin Valley, California," Journal of hydrology (Amsterdam) 569, 569 (February). Elsevier B.V: 462–69. doi:10.1016/j.jhydrol.2018.11.077.

"Land Subsidence Due to Ground-Water Withdrawal Tulare-Wasco Area California GEOLOGICAL SURVEY PROFESSIONAL PAPER 437-B Prepared in Cooperation with the California Department of Water Resources." n.d.

Lees, M., Knight, R. and Smith, R.. 2021. "Modeling 65 Years of Land Subsidence in California's San Joaquin Valley." Research Square Platform LLC. doi:10.21203/rs.3.rs-609832/v1.

LeSage, J. P., and Fischer, M. M. 2008. Spatial growth regressions: Model specification, estimation and interpretation. Spatial Economic Analysis, 3 (3), 275–304.

Levy, M.C., Neely, W.R., Borsa, A.A., and Burney, J.A. 2020. "Fine-Scale Spatiotemporal Variation in Subsidence across California's San Joaquin Valley Explained by Groundwater Demand." Environmental Research Letters 15. IOP Publishing. doi:10.1088/1748-9326/abb55c.

Liu, D., Zhongrui F., Fu, Q., Li, M., Faiz, M.A., Ali, S., Li, T., Zhang, L., and Khanm, M.I. 2020. "Random Forest Regression Evaluation Model of Regional Flood Disaster Resilience Based on the Whale Optimization Algorithm," Journal of cleaner production 250, 250 (March). Elsevier Ltd: 119468. doi:10.1016/j.jclepro.2019.119468.

Liu, Z., Liu, P.W., Massoud, E., Farr, T.G., Lundgren, P., and Famiglietti, J.S. 2019. "Monitoring Groundwater Change in California's Central Valley Using Sentinel-1 and GRACE Observations," Geosciences 9, 9 (10). Basel: MDPI AG: 436. doi:10.3390/geosciences9100436.

Matthews, S. A., and Yang, T. C. 2012. Mapping the results of local statistics: Using geographically weighted regression. Demographic research, 26, 151-166.

McPherson, J.G., and Miller, D.D., 1990. Depositional Settings and Reservoir Characteristics of the Plio-Pleistocene Tulare Formation, South Belridge Field, San Joaquin Valley, California. Edited by Jonathon G. Kuespert and Stephen A. Reid. Structure, Stratigraphy and Hydrocarbon Occurrences of the San Joaquin Basin, California. The Pacific Section American Association of Petroleum Geologist. doi:10.32375/1990-GB65.17.

Miller, R. E. Green, J. H. and Davis, G. H. 1971. Geology of the compacting deposits in the Los Banos-Kettleman City subsidence area, California: U.S. Geological Survey Professional Paper 497-E, 46 p.

Miller, M.M., Shirzaei, M., 2015. Spatiotemporal characterization of land subsidence and uplift in Phoenix using InSAR time series and wavelet transforms. J. Geophys. Res. Solid Earth 120, 5822–5842.

Mitchell, A. *The ESRI Guide to GIS Analysis,* Volume 1. 2nd Ed. ESRI Press, 2020

Mitchell, A. and Griffin, L.S. *The ESRI Guide to GIS Analysis,* Volume 2. 2nd Ed. ESRI Press, 2021.

Muarry, C. J., Low, D.R., Graham, S.A., Martinez, P.A. 1996. Statistical Analysis of Bed-Thickness Patterns in A Turbidite Section from the Great Valley Sequence, Cache Creek, Northern California: Journal of Sedimentary Research, v. 66, no. 5, p. 900-908.

Murray, K.D., and Lohman, R.B. 2018. "Short-Lived Pause in Central California Subsidence after Heavy Winter Precipitation of 2017," Science Advances 4, 4 (8). United States: American Association for the Advancement of Science: eaar8144. doi:10.1126/sciadv.aar8144.

Narasimhan, T. N. 2008. "Comment on 'A Method to Estimate Groundwater Depletion from Confining Layers' by L. F. Konikow and C. E. Neuzil," Water resources research 44, 44 (6). American Geophysical Union: W06605-n/a. doi:10.1029/2008WR006863.

Nelder, J. A. and Wedderburn, R. W. M. (1972) Generalized linear models. J. R. Statist. Soc. A, 135, 370 - 384.

Neuendorf, K.K.E., Mehl, J.P., Jr., and Jackson, J.A., eds., 2011, Glossary of Geology: Alexandria, Virginia, American Geological Institute, p. 642

NOAA. "San Joaquin River Basin." Accessed November 30, 2022. https://www.fisheries.noaa.gov/west-coast/habitat-conservation/san-joaquin-river-basin

Page, B. M. 1966. Geology of the Coast Range of California, chapter 6 in Geology of Northern California: California Division of Mines Bulletin 190, p. 255-276.

Page, R. W. 1973. Base of fresh ground water (approximately 3,000 micromhos) in the San Joaquin Valley, California: U.S. Geological Survey Hydrologic Investigations Atlas HA-489.

Page, R. W. 1983. "Data on Depths to the Upper Mya Zone of the San Joaquin Formation in the Kettleman City Area, San Joaquin Valley, California." US Geological Survey. doi:10.3133/ofr81699.

Poland, J. F., Lofgren, B. E., Ireland, R. L. & Pugh, R. G. 1972. Land Subsidence in the San Joaquin Valley, California, As of 1972. 87

Sneed, M. and Brandt, J. 2013. Detection and Measurement of Land Subsidence Using Global Positioning System Surveying and Interferometric Synthetic Aperture Rader, Coachella Valley, California, 1996-2005: U.S. Geological Survey Scientific Investigations Report 2007-5251, 30 p.

Sneed, M. 2018. In: Alley, W.M., ed., *Groundwater: State of the Science and Practice*, Westerville, Ohio, National Groundwater Association, p. 58-62.

Stapleton, J.H. Linear Statistical Models. Germany: Wiley, 2009.

Sustainable Groundwater Management Act of 2014, Water Code – WAT, Division 6, Part 2.74, § 10720 -10737.8 Stats. Ch. 346, Sec. 3 (2014) Sustainable Groundwater Management Act (SGMA) Data Viewer. Accessed March 5, 2022: https://sgma.water.ca.gov/webgis/?appid=SGMADataViewer#currentconditions

Talling, P. J. 2001. On the frequency distribution of turbidite thickness: Sedimentology, v. 48, p. 1297-1329.

Thapa RB, Estoque RC. 2012. Geographically Weighted Regression in Geospatial Analysis. Y Murayama (Ed.) Progress in Geospatial Analysis (pp. 85-96). Tokyo: Springer. Accessed directly through SpringerLink: http://link.springer.com/chapter/10.1007/978-4-431-54000-7_6

Tuson, M., Yap, M., Kok, M.R., Boruff, B., Murray, K., Vickery, A., Turlach, B.A., and Whyatt, D. "Overcoming inefficiencies arising due to the impact of the modifiable areal unit problem on single-aggregation disease maps." International Journal of Health Geographics 19. 40 (2020).

Twiss, R.J. and Moores, E.M. 2001. Structural Geology. W.H. Freeman and Company, New York, 532.

USGS. 2022. "Areas of Land Subsidence in California." Accessed March 5, 2022: https://ca.water.usgs.gov/land_subsidence/california-subsidence-areas.html

USGS. 2022. "Delta-Mendota Canal: Evaluation of Gorundwater Conditions & Land Subsidence." Accessed August 24, 2022: https://ca.water.usgs.gov/projects/central-valley/delta-mendota-canal.html

USGS. 2022. "The Central Valley: Tulare Basin." Accessed June 6, 2022: https://ca.water.usgs.gov/projects/central-valley/tulare-basin.html

Walton, W. C. 1979. "Progress in Analytical Groundwater Modeling," Journal of Hydrology 43, 43 (1): 149–59. doi:10.1016/0022-1694(79)90170-7.

Williamson, A.K., Prudic, D.E., and Swain, L.A. 1989. Ground-water flow in the Central Valley, California: U.S. Geological Survey Professional Paper 1401-D.

Yue, J., Cheng, W., and Fang, L.2009. "The Study on Mathematical Model of Urban Land Subsidence Based on Statistical Analysis," 2009 International Conference on Management and Service Science, 2009, pp. 1-4, doi: 10.1109/ICMSS.2009.5305660.

Appendix A – Summary of OLS Regression Variable Coefficients

**Indicates a statistically significant relationship*

Variable	Coefficient	Robust SE	Robust t	Robust Pr	VIF
2015 Intercept	0.029543	0.01447	2.041646	0.041269*	--------
Total Well Depth	-0.000002	0.00001	-0.193992	0.84619	3.797276
Well Completion Length	-0.000012	0.000014	-0.799482	0.424066	2.724941
2015 Groundwater Level	0.000133	0.000019	7.178395	0.000000*	1.389416
Depth to Corcoran Clay	-0.000218	0.000015	-14.248787	0.000000*	1.865914
Corcoran Clay Thickness	-0.000784	0.000161	-4.87007	0.000002*	1.774136
% Fine Grain Material	-0.002617	0.000202	-12.971017	0.000000*	1.0709
Up vs. Lower Tulare	0.049965	0.009337	5.351401	0.000000*	2.105291
2016 Intercept	-0.003831	0.012478	-0.307051	0.758835	--------
Total Well Depth	0.00001	0.000009	1.008106	0.313466	3.900945
Well Completion Length	-0.000014	0.000014	-0.97842	0.327925	2.673596
2016 Groundwater Level	0.000373	0.000017	22.223439	0.000000*	1.392284
Depth to Corcoran Clay	-0.001942	0.000175	-11.079092	0.000000*	1.065782
Corcoran Clay Thickness	-0.00035	0.000014	-24.690774	0.000000*	1.72856
% Fine Grain Material	-0.000534	0.000117	-4.560865	0.000007*	1.208554
Up vs. Lower Tulare	0.040189	0.007846	5.121859	0.000001*	2.152318
2017 Intercept	-0.160512	0.010603	-15.138003	0.000000*	--------
Total Well Depth	0.000024	0.000007	3.584269	0.000358*	3.726871
Well Completion Length	0.000022	0.000011	2.033164	0.042121*	2.557095
2017 Groundwater Level	-0.000156	0.000015	-10.254633	0.000000*	1.355889
Depth to Corcoran Clay	0.001209	0.000152	7.967612	0.000000*	1.093304
Corcoran Clay Thickness	0.000375	0.000013	29.571774	0.000000*	1.664723
% Fine Grain Material	-0.000626	0.000077	-8.096861	0.000000*	1.152584
Up vs. Lower Tulare	0.021195	0.006339	3.343754	0.000854*	2.179722
2018 Intercept	0.021678	0.009533	2.273947	0.023021*	--------
Total Well Depth	0.000014	0.000006	2.457641	0.014027*	3.787331
Well Completion Length	-0.000015	0.000009	-1.700012	0.089239	2.625466
2018 Groundwater Level	0.000152	0.000012	12.440421	0.000000*	1.355912
% Fine Grain Material	-0.001437	0.000135	-10.613741	0.000000*	1.097181
Depth to Corcoran Clay	-0.000157	0.000009	-16.639823	0.000000*	1.659258
Corcoran Clay Thickness	-0.000363	0.000094	-3.878788	0.000117*	1.201381
Up vs. Lower Tulare	0.02254	0.005606	4.02056	0.000067*	2.184942
2019 Intercept	-0.054736	0.006505	-8.414983	0.000000*	--------
Total Well Depth	0.000037	0.000007	5.638191	0.000000*	3.484539
Well Completion Length	-0.000032	0.000011	-2.945419	0.003276*	2.870938

2019 Groundwater Level	0.000071	0.00001	7.032167	0.000000*	1.509999
% Fine Grain Material	-0.000589	0.000098	-5.990892	0.000000*	1.018384
Depth to Corcoran Clay	-0.000003	0.000011	-0.273684	0.784366	1.232173
Corcoran Clay Thickness	0.000115	0.000054	2.119753	0.034153*	1.320275
Up vs. Lower Tulare	0.001292	0.0045	0.287129	0.774056	1.966936
2020 Intercept	0.111662	0.0137	8.150752	0.000000*	--------
Total Well Depth	-0.00004	0.000012	-3.463482	0.000564*	3.245993
Well Completion Length	0.00001	0.000018	0.558697	0.576459	2.885067
2020 Groundwater Level	0.000131	0.000018	7.268299	0.000000*	1.454729
% Fine Grain Material	-0.001483	0.000156	-9.506363	0.000000*	1.053744
Depth to Corcoran Clay	-0.000384	0.000023	-16.832402	0.000000*	1.335624
Corcoran Clay Thickness	-0.000491	0.000124	-3.966258	0.000085*	1.216716
Up vs. Lower Tulare	0.010778	0.00725	1.486724	0.137312	1.958827
2021 Intercept	0.133876	0.014829	9.028173	0.000000*	--------
Total Well Depth	-0.000022	0.000013	-1.688206	0.091584	3.230477
Well Completion Length	0.000004	0.000022	0.186596	0.851995	2.808253
2021 Groundwater Level	-0.000023	0.00002	-1.187614	0.235168	1.630424
% Fine Grain Material	-0.00258	0.000181	-14.260905	0.000000*	1.066257
Depth to Corcoran Clay	-0.000484	0.000024	-20.26942	0.000000*	1.711205
Corcoran Clay Thickness	-0.000115	0.000152	-0.755654	0.449962	1.777817
Up vs. Lower Tulare	0.013586	0.008655	1.569676	0.11671	1.917885

Appendix B - R Code for EDA

```
# Calling Spatial Data in R (sp)
# SSCI 594b EDA Code

# Subsidence and Groundwater in the San Joaquin Valley CA

# call packages from library
library(sp)
library(rgdal)

#load R Excel reader package
library(readxl)

# call data in Excel and name vector / variable
SJB <- read_excel("/Volumes/GoogleDrive/My
Drive/USC/MSGIST/Thesis/Datasets/2017TESTforMorans.xlsx")

# check field headers
head(SJB)

#check the number of rows
nrow(SJB)

# check the number of columns with recorded data
ncol(SJB)

# check variable names of each data field of each layer
names(SJB)

#call dataset into sp package
str(SJB)
class(SJB)
data = data.frame(SJB)

# set XY data to dataframe
coords = data.frame(
 x=SJB$LONGITUDE,
 y=SJB$LATITUDE
)

# plot cooridnates from dataframe
plot(coords,pch=21, bg="lightblue", xlab = "Longitude", ylab = "Latitude", main = "Lat & Long
Locations of Water Wells - SJV")
grid(nx= NULL,
   ny= NULL,
```

```
   lty = 2, col= "dark grey", lwd =1)

coordinates(data) = cbind(coords$x, coords$y)

#get summary stats
summary(SJB)

#### Plot Histograms

hist(SJB$SUBSIDENCE_RATE, xlab = "Subsidence (ft)", main = "Distribution of Land
Subsidence (2017)", xlim = c(-1, 0.5), ylim = c(0, 1000), breaks = 15)
hist(SJB$PERCENT_FINE, xlab = "Percent Fine Grain Sediment (%)", main = "Distribution of
Percent Fine Grain Sediment (2021)", xlim = c(0, 100), ylim = c(0, 500), breaks = 15)
hist(SJB$PERCENT_COARSE,xlab = "Percent Coarse Grain Sediment (%)", main =
"Distribution of Percent Coarse Grain Sediment (2021)", xlim = c(0, 100), ylim = c(0, 500),
breaks = 15)
hist(SJB$COMPL_LENGTH,xlab = "Completion Lengths (ft)", main = "Distribution of
Completion Lengths (2021)", xlim = c(0, 2500), ylim = c(0, 2500), breaks = 25)
hist(SJB$PERF_TOP_DEPTH, xlab = "Depth of Top Perforation (ft bgs)", main = "Distribution
of Top Perforation Depths (2021)", xlim = c(0, 2000), ylim = c(0, 2000), breaks = 25)
hist(SJB$PERF_BASE_DEPTH, xlab = "Depth of Base Perforation (ft bgs)", main =
"Distribution of Base Perforation Depths (2021)", xlim = c(0, 3000), ylim = c(0, 2000), breaks =
25)
hist(SJB$WELL_DEPTH, xlab = "Total Depth (ft bgs)", main = "Distribution of Total Well
Depths (2021)", xlim = c(0, 3000), ylim = c(0, 2000), breaks = 20)
hist(SJB$GWLEVEL_2017, xlab = "Groundwater Depth (ft bgs)", main = "Distribution of
Groundwater Depth (2021)", xlim = c(-10, 1500), ylim = c(0, 1500), breaks = 15)
hist(SJB$GW_LEVEL_CHANGE, xlab = "Groundwater Level Change", main = "Distribution of
Groundwater Level Change (2021)", xlim = c(-50, 50), ylim = c(0, 3500), breaks = 10)
hist(SJB$UP_LOW_TULARE, xlab = "Upper:Lower Tulare", main = "Distribution of Upper vs.
Lower Tulare Well Completions (2021)", xlim = c(0, 1), ylim = c(0, 3500), breaks = 10)

hist(SJB$DEPTH_COR_CLAY, xlab = "Depth (ft bgs)" , main = "Distribution of Top Corcoran
Clay Depths (ft bgs)", xlim = c(0, 1000), ylim = c(0, 500), breaks = 15 )
hist(SJB$COR_CLAY_THICK,xlab = "Thickness (ft)" , main = "Distribution of Corcoran Clay
Thickness", xlim = c(0, 150), ylim = c(0, 700), breaks = 15 )

##hist(SJB$SUBSIDENCE_RATE,xlab = "Land Subsidence (cm)", main = "Distribution of
Land Subsidence (2021)", xlim = c(-1.5, 0.5), ylim = c(0, 1500), breaks = 15)

### Jarque Bera Test for Normality ###

# call package from library

library(tseries)
```

```
# run normality test
# p-value < 0.05 = not normal
# p-value > 0.05 = normal distribution
dataset <- rnorm(SJB$SUBSIDENCE_RATE) # null
jarque.bera.test(dataset)

dataset <- rnorm(SJB$PERCENT_FINE) # null
jarque.bera.test(dataset)

dataset <- rnorm(SJB$PERCENT_COARSE) # null
jarque.bera.test(dataset)

dataset <- rnorm(SJB$COMPL_LENGTH) # null
jarque.bera.test(dataset)

dataset <- rnorm(SJB$PERF_TOP_DEPTH) # null
jarque.bera.test(dataset)

dataset <- rnorm(SJB$PERF_BASE_DEPTH) # null
jarque.bera.test(dataset)

dataset <- rnorm(SJB$WELL_DEPTH) # null
jarque.bera.test(dataset)

dataset <- rnorm(SJB$GWLEVEL_2017) # null
jarque.bera.test(dataset)

dataset <- rnorm(SJB$GW_LEVEL_CHANGE) # null
jarque.bera.test(dataset)

dataset <- rnorm(SJB$UP_LOW_TULARE) # null
jarque.bera.test(dataset)

dataset <- rnorm(SJB$COR_CLAY_THICK) # null
jarque.bera.test(dataset)

dataset <- rnorm(SJB$DEPTH_COR_CLAY) # null
jarque.bera.test(dataset)

#### KS test for normality ####

# Note that p<0.5 means the data are not normally distributed
# Note that p>0.5 means the data are normally distributed
ks.test(SJB$PERCENT_FINE, 'pnorm')

# ks.test(SJB$PERCENT_COARSE, 'pnorm')
```

```
# ks.test(SJB$GWLEVEL_2021, 'pnorm')

### Log Transform Non-Normal Distributions ####

# Check existing data frame
data = data.frame(SJB)
data.frame(SJB)

# Add named column to data frame and perform log transformation
SJB$logsubsidence=log(SJB$Subsidence)

#check resulting data field and frame
data.frame(SJB)

### Rerun histogram of log transformed data fields from data frame
hist(SJB$logsubsidence, xlab = "Subsidence (ft)", main = "Histogram of Log Transformed
Subsidence - SJ Basin", xlim = c(-1.5, 0), ylim = c(0, 3000), breaks = 13)

# Log Transformation of Groundwater Depth data

# Add named column to data frame and perform log transformation
SJB$logdepth=log(SJB$GSE)

#check resulting data field and frame
data.frame(SJB)

### Rerun histogram of log transformed data fields from data frame
hist(SJB$logdepth, xlab = "Groundwater Depth (ftbgs)", main = "Histogram of Log Transformed
Groundwater Depth - SJ Basin", xlim = c(0, 10), ylim = c(0, 3000), breaks = 20)

### Make scatter plots ###

#plot data for first run in preparation for background "underlay" color
plot(SJB$GWLEVEL_2017, SJB$SUBSIDENCE_RATE, main = "Subsidence / Groundwater
Depth (2017)", xlab="Groundwater Depth (ft bgs)", ylab="Subsidence (cm)", pch=21, bg="light
blue")

# set background color "underlay"
rect(par("usr")[1], par("usr")[3],
  par("usr")[2], par("usr")[4],
  col = "gray96")

# add a new plot area in current graph
par(new= TRUE)

# plot selected data in scatterplot
```

```
grid(nx= NULL,
  ny= NULL,
  lty = 2, col= "dark grey", lwd =1)
par(new= TRUE)
plot(SJB$GWLEVEL_2017, SJB$SUBSIDENCE_RATE, main = "Subsidence / Groundwater
Depth (2017)", xlab="Groundwater Depth (ft bgs)", ylab="Subsidence (cm)", pch=21, bg="light
blue")
### Boxplots for Groundwater wells ###

# set parameters for alternating boxplot colors by y~group
# las = 2 changes the label direction while cex.axis = 0.5 changes the axis labels' font size
boxplot(SJB$GWLEVEL_2017~SJB$BASIN_NAME,
data=data.frame(SJB$GWLEVEL_2017), col=(c("lightgreen", "lightblue")),
main="Groundwater Depth by Basin (2017)", ylab="Groundwater Depth (ft bgs)", xlab="Basin",
las=2, cex.axis=0.5)

# set background color "underlay"
rect(par("usr")[1], par("usr")[3],
  par("usr")[2], par("usr")[4],
  col = "gray96")

# add a new plot area in current graph
par(new= TRUE)

# plot selected data in scatterplot
grid(nx= NA, # no grid lines
  ny=NULL, # default grid line per label
  lty = 2, col= "dark grey", lwd =1)

par(new= TRUE)

# set parameters for alternating boxplot colors by y~group
boxplot(SJB$GWLEVEL_2017~SJB$BASIN_NAME,
data=data.frame(SJB$GWLEVEL_2017), col=(c("lightgreen", "lightblue")),
main="Groundwater Depth by Basin (2017)", ylab="Groundwater Depth (ft bgs)", xlab="Basin",
las=2, cex.axis=0.5)

### Boxplots for Subsidence Monitoring Network ###

# set parameters for alternating boxplot colors by y~group
# las = 2 changes the label direction while cex.axis = 0.5 changes the axis labels' font size
boxplot(SJB$GW_LEVEL_CHANGE~SJB$BASIN_NAME,
data=data.frame(SJB$GW_LEVEL_CHANGE), col=(c("lightcoral", "lightgoldenrod1")),
main="Groundwater Level Change by Basin (2017)", ylab="Groundwater Level Change (ft)",
xlab="Basin", las=2, cex.axis=0.75)
```

```
# set background color "underlay"
rect(par("usr")[1], par("usr")[3],
  par("usr")[2], par("usr")[4],
  col = "gray96")
# add a new plot area in current graph
par(new= TRUE)

# plot selected data in scatterplot
grid(nx= NA, # no grid lines
  ny=NULL, # default grid line per label
  lty = 2, col= "dark grey", lwd =1)

par(new= TRUE)

boxplot(SJB$GW_LEVEL_CHANGE~SJB$BASIN_NAME,
data=data.frame(SJB$GW_LEVEL_CHANGE), col=(c("lightcoral", "lightgoldenrod1")),
main="Groundwater Level Change by Basin (2017)", ylab="Groundwater Level Change (ft)",
xlab="Basin", las=2, cex.axis=0.75)

### Plot 3D Scatter Plot ####

# call library
library(rgl)
# plot X, Y, Z data
# By default, plot3d() uses square points, which do not appear properly when saving to a PDF.
# For improved appearance, the example above uses type="s" for spherical points, made them
smaller with size=0.75, and turned off the 3D lighting with lit=FALSE (otherwise they look like
shiny spheres)

plot3d(SJB$LONGITUDE, SJB$LATITUDE, SJB$GWLEVEL_2015,
   xlab = "Longitude", ylab = "Latitude", zlab = "Groundwater Depth (ft)",
   type = "s", size = 0.25, col="lightblue", lit = FALSE)

# plot x, y, z subsidence data
plot3d(SJB$LONGITUDE, SJB$LATITUDE, SJB$SUBSIDENCE_RATE,
   xlab = "Longitude", ylab = "Latitude", zlab = "Subsidence (ft)",
   type = "s", size = 0.25, col="red", lit = FALSE)

##attempt to add segments to each water depth -- makes things too messy
##segments3d(interleave(SJGW$LONGITUDE, SJGW$LONGITUDE),
##     interleave(SJGW$LATITUDE, SJGW$LATITUDE),
##    interleave(SJGW$GSE), min(SJGW$GSE),
##   alpha = 0.4, col= "blue")
```

Appendix C - R Code for ESDA

```
#### Moran's I for Spatial Autocorrelation ###
### Spatially Lagged Model ###

### Set Neighborhood for Large Pointset ####

# call libraries
library(rgdal)
library(spdep)
library(readxl)
library(ape)

# read and alias dataset from Excel
SJB <-
read_excel("/Volumes/GoogleDrive/MyDrive/USC/MSGIST/Thesis/Datasets/2017TESTforMor
ans.xlsx")
# Check data field titles and categories
head(SJB)

# Make data a spatial dataframe
coordinates(SJB) <- ~ LONGITUDE + LATITUDE

# set inverse distance weights matrix
well.dists <-as.matrix(dist(cbind(SJB$LONGITUDE, SJB$LATITUDE)))

# 1 divided by values in the matrix
well.dist.inv <- 1/well.dists

# create a matrix where each-off diagonal [i,j] is equal to 1/distance between point i and point j
diag(well.dist.inv) <- 0
well.dist.inv[1:5, 1:5]

# Calculate Moran's I (note that I am not too sure if this step is needed here)
Moran.I(SJB$SUBSIDENCE_RATE, well.dist.inv)

# Use K nearest as a criteria for non-polygon vectors (ie points)

knea <- knearneigh(coordinates(SJB), longlat = TRUE)
neib <- knn2nb(knea)

# Print output of the nearest neighbors
neib
par = (mar=c(0,0,0,0))
plot(SJB, border="grey")
plot(neib, coordinates(SJB), add=TRUE, col="red")
```

```
### Global Moran's I ###
# assign weights as listed object
#create a spatial weights using nb2listw() using the default option of row standardization
(style="W") and binary weights (style ="B")
SJBW <- nb2listw(neib, style = "W", zero.policy = TRUE)
lw <- nb2listw(neib, style = "B")
head(SJB)

# Assess the weight of the first polygon's neighbors type
lw$weights

# Compute weighted neighbor mean subsidence values
# not needed to run the moran or moran.test
# Compute the average neighbor subsidence value for each polygon or spatially lagged values
inc.lag <- lag.listw(lw, SJB$SUBSIDENCE_RATE)
inc.lag


# Plot relationship between spatially lagged neighbors and subsidence
# Fitted line is part of the OLS model

plot(inc.lag ~ SJB$SUBSIDENCE_RATE, pch=16, asp=1)
M1 <- lm(inc.lag ~SJB$SUBSIDENCE_RATE)


# plot regression line (OLS)
abline(M1, col="red")

# Print the slope of the regression line (or variance) for Moran's I coefficient
coef(M1) [2]

# Compute the Moran's I statistic
I <- moran(SJB$SUBSIDENCE_RATE, lw, length(neib), Szero(lw)) [1]
I
# Perform a hypothesis test
# testing if the subsidence values are randomly distributed across each basin following a
completely random process
# this can be tested with an anlytical method and stochastic method: Monte Carlo method

# Analytical Method
# Note thtat -1 is perfect clustering of dissimliar values (ie perfect dispersion)
# Note that 0 is no spatial autocorrelation
# Note that +1 indicates perfect clustering of similar values
moran.test(SJB$SUBSIDENCE_RATE, lw, alternative = "greater")

# Moran's I plot of Subsidence
```

```
moran.plot(SJB$SUBSIDENCE_RATE,listw=nb2listw(neib, style="C"))


### Spatial Lag Model ###

#be sure to open library for spatial regression before running the script below the library call
library(spatialreg)

#Here if you look at the Coefficients part of the result, all of my 6 regression coefficients are tested statistically significant (against 0).

#For example, my feemp coefficient is -0.5187 with standard error 0.0114. The Z-test statistic (z-score) is -45.5902 and the p-value is smaller than 2.2E-16.

#This tells us that a 1% increase in the female employment rate is associate with a 0.5187% decrease in the poverty rate when all other explanatory variables are held constant.

m3_lag <- lagsarlm(SJB$SUBSIDENCE_RATE ~ SJB$WELL_DEPTH +
SJB$PERF_TOP_DEPTH + SJB$PERF_BASE_DEPTH + SJB$COMPL_LENGTH +
SJB$GWLEVEL_2017 + SJB$PERCENT_FINE + SJB$PERCENT_COARSE +
SJB$DEPTH_COR_CLAY + SJB$COR_CLAY_THICK + SJB$GW_LEVEL_CHANGE +
SJB$UP_LOW_TULARE, data = SJB, listw = lw, type = "lag", zero.policy = TRUE)

summary(m3_lag, correlation=FALSE)
#Extract estimated regression coefficients and the corresponding 95% confidence intervals to a table of three columns, one for the estimated regression coefficients
# and the other two for the lower and upper limits of the 95% CI.
cbind(coefest= coef(m3_lag), confint(m3_lag))

# plot residuals against fitted responses
plot(m3_lag$fitted.values, m3_lag$residuals, xlab="Fitted Values", ylab="Residuals", main
="Residuals vs Fitted", cex=0.1)
ab <- lm(m3_lag$fitted.values ~m3_lag$residuals)

abline(ab, lty=2, col="red")
```

Appendix D - R Code for Spatial Regression Models

```
### OLS Model ###


#load R Excel reader package
library(readxl)

# call data in Excel and name vector / variable
SJB <- read_excel("/Volumes/GoogleDrive/My
Drive/USC/MSGIST/Thesis/Datasets/2017TESTforMorans.xlsx")

# check field headers
head(SJB)

# run lm test for Groundwater Level Change and Subsidence Rate
OLS = lm(as.numeric(SJB$GW_LEVEL_CHANGE)~as.numeric(SJB$SUBSIDENCE_RATE),
data=SJB)

# plot OLS model
plot(OLS)

# get pvalues and R2 values
summary(OLS)

# print p-value
summary(OLS)$coefficients[,4]

# print r2 value
summary(OLS)$r.squared

summary(OLS)$adj.r.squared

#### GWLEVEL_YEAR and SUBSIDENCE RATE OLS and Plots

# lm test for groundwater level
OLS = lm(SJB$GWLEVEL_2017~SJB$SUBSIDENCE_RATE, data=SJB)
plot(OLS)

# get pvalues and R2 values
summary(OLS)

# print p-value
summary(OLS)$coefficients[,4]
#print r2 value
summary(OLS)$r.squared
```

```
#### Percent Fine and SUBSIDENCE RATE OLS and Plots

# lm test for groundwater level
OLS = lm(SJB$PERCENT_FINE~SJB$SUBSIDENCE_RATE, data=SJB)
plot(OLS)

# get pvalues and R2 values
summary(OLS)

# print p-value
summary(OLS)$coefficients[,4]

#print r2 value
summary(OLS)$r.squared

#### Percent Coarse and SUBSIDENCE RATE OLS and Plots

# lm test for groundwater level
OLS = lm(SJB$PERCENT_COARSE~SJB$SUBSIDENCE_RATE, data=SJB)
plot(OLS)

# get pvalues and R2 values
summary(OLS)

# print p-value
summary(OLS)$coefficients[,4]

#print r2 value
summary(OLS)$r.squared

#### Completion Length and SUBSIDENCE RATE OLS and Plots

# lm test for Completion length
OLS = lm(as.numeric(SJB$COMPL_LENGTH)~as.numeric(SJB$SUBSIDENCE_RATE),
data=SJB)
plot(OLS)

# get pvalues and R2 values
summary(OLS)

# print p-value
summary(OLS)$coefficients[,4]
#print r2 value
summary(OLS)$r.squared
```

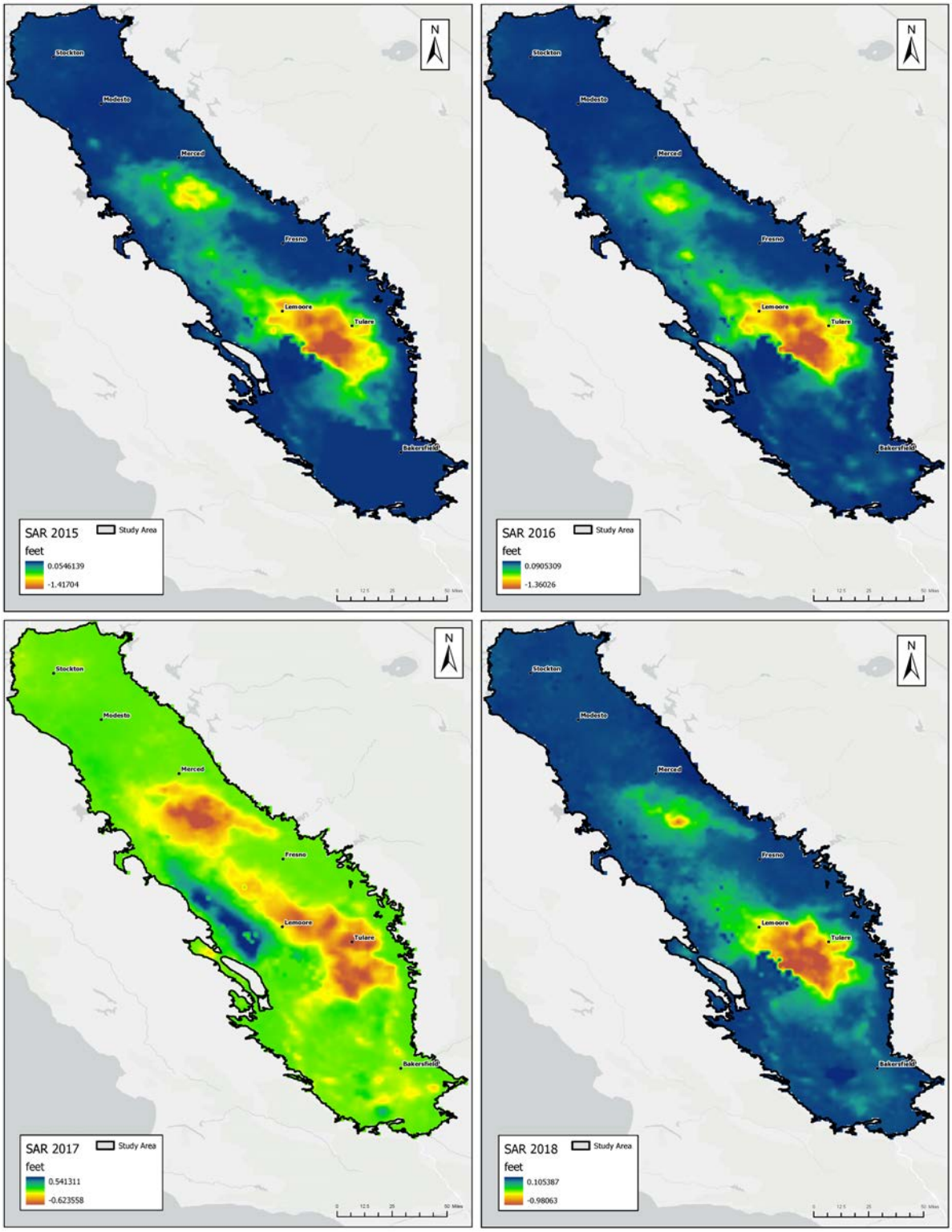
```
#### Perf Top Depth and SUBSIDENCE RATE OLS and Plots

# lm test for top perf depth
OLS = lm(as.numeric(SJB$PERF_TOP_DEPTH)~as.numeric(SJB$SUBSIDENCE_RATE),
data=SJB)
plot(OLS)

# get pvalues and R2 values
summary(OLS)

# print p-value
summary(OLS)$coefficients[,4]
#print r2 value
summary(OLS)$r.squared

#### Perf Base Depth and SUBSIDENCE RATE OLS and Plots

# lm test for perf base depth
OLS = lm(as.numeric(SJB$PERF_BASE_DEPTH)~as.numeric(SJB$SUBSIDENCE_RATE),
data=SJB)
plot(OLS)

# get pvalues and R2 values
summary(OLS)

# print p-value
summary(OLS)$coefficients[,4]
#print r2 value
summary(OLS)$r.squared

#### Well Depth and SUBSIDENCE RATE OLS and Plots

# lm test for well total depth
OLS = lm(SJB$WELL_DEPTH~as.numeric(SJB$SUBSIDENCE_RATE), data=SJB)
plot(OLS)

# get pvalues and R2 values
summary(OLS)

# print p-value
summary(OLS)$coefficients[,4]
#print r2 value
summary(OLS)$r.squared

#### Upper vs Lower Tulare Completion and SUBSIDENCE RATE OLS and Plots
```

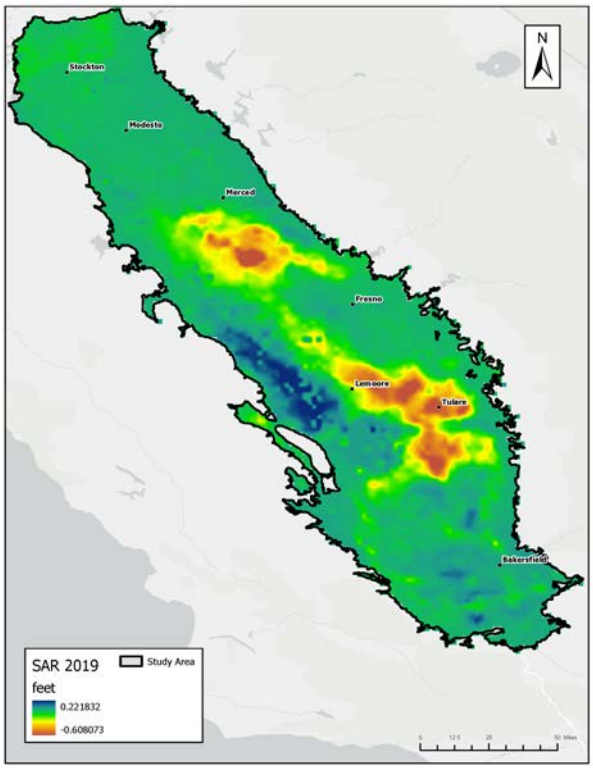
```
# lm test for upper vs lower Tulare
OLS = lm(SJB$UP_LOW_TULARE~as.numeric(SJB$SUBSIDENCE_RATE), data=SJB)
plot(OLS)

# get pvalues and R2 values
summary(OLS)

# print p-value
summary(OLS)$coefficients[,4]
#print r2 value
summary(OLS)$r.squared


#### COR_CLAY_THICK and SUBSIDENCE RATE OLS and Plots

# lm test for upper vs lower Tulare
OLS = lm(SJB$COR_CLAY_THICK~as.numeric(SJB$SUBSIDENCE_RATE), data=SJB)
plot(OLS)

# get pvalues and R2 values
summary(OLS)

# print p-value
summary(OLS)$coefficients[,4]
#print r2 value
summary(OLS)$r.squared


#### DEPTH_COR_CLAY and SUBSIDENCE RATE OLS and Plots

# lm test for upper vs lower Tulare
OLS = lm(SJB$DEPTH_COR_CLAY~as.numeric(SJB$SUBSIDENCE_RATE), data=SJB)
plot(OLS)

# get pvalues and R2 values
summary(OLS)

# print p-value
summary(OLS)$coefficients[,4]
#print r2 value
summary(OLS)$r.squared

#### Test for Multicolinearity ####

#call library
library(car)
```

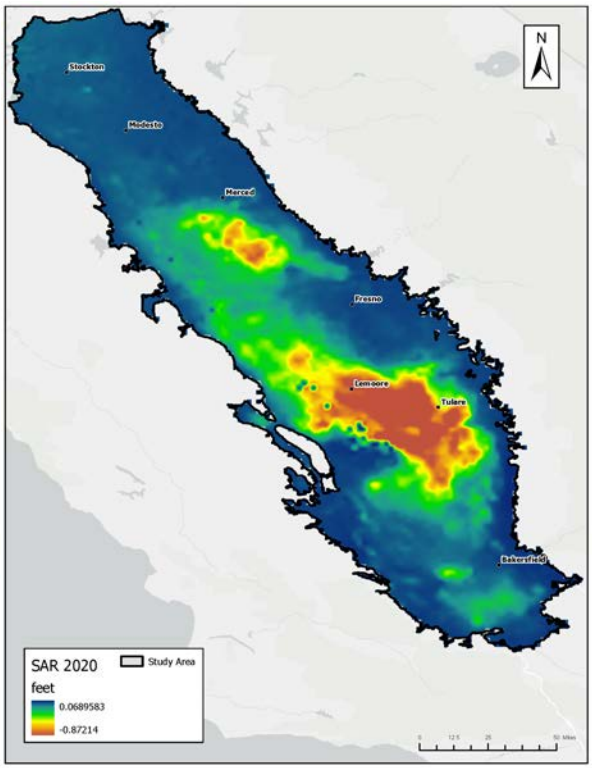
```
#Fit the regression model (looking at values close to zero to see if it makes sense to include the variable)
lm(SJB$SUBSIDENCE_RATE ~ SJB$WELL_DEPTH + SJB$PERF_TOP_DEPTH + SJB$PERF_BASE_DEPTH + SJB$COMPL_LENGTH + SJB$GWLEVEL_2017 + SJB$PERCENT_FINE + SJB$PERCENT_COARSE + SJB$DEPTH_COR_CLAY + SJB$COR_CLAY_THICK + SJB$GW_LEVEL_CHANGE + SJB$UP_LOW_TULARE, data=SJB)

# check for multicollinearity

#place variables in data frame
df <- data.frame(as.numeric(SJB$WELL_DEPTH), as.numeric(SJB$PERF_TOP_DEPTH), as.numeric(SJB$PERF_BASE_DEPTH), as.numeric(SJB$COMPL_LENGTH), as.numeric(SJB$GWLEVEL_2017), as.numeric(SJB$PERCENT_FINE), as.numeric(SJB$PERCENT_COARSE), as.numeric(SJB$DEPTH_COR_CLAY), as.numeric(SJB$COR_CLAY_THICK), as.numeric(SJB$GW_LEVEL_CHANGE), as.numeric(SJB$UP_LOW_TULARE), SJB$SUBSIDENCE_RATE)

#create correlation matrix for data frame
cor(df)

#calculate the variance inflation factor (VIF) for each predictor variable in the model
# vif was run for each variable until no aliased variables (ie multicollinearity) were present
vif(lm(SJB$SUBSIDENCE_RATE~SJB$WELL_DEPTH + SJB$COMPL_LENGTH + SJB$GWLEVEL_2017 + SJB$PERCENT_FINE + SJB$DEPTH_COR_CLAY + SJB$COR_CLAY_THICK + SJB$GW_LEVEL_CHANGE + SJB$UP_LOW_TULARE))

###### Regression Model Selection by AIC #####

#Fit the regression model (looking at values close to zero to see if it makes sense to include the variable)

m1=lm(SJB$SUBSIDENCE_RATE ~ SJB$WELL_DEPTH + SJB$PERF_TOP_DEPTH + SJB$PERF_BASE_DEPTH + SJB$COMPL_LENGTH + SJB$GWLEVEL_2017 + SJB$PERCENT_FINE + SJB$PERCENT_COARSE + SJB$DEPTH_COR_CLAY + SJB$COR_CLAY_THICK + SJB$GW_LEVEL_CHANGE + SJB$UP_LOW_TULARE, data=SJB)
summary(m1)

cbind(coefest=coef(m1), confint(m1))

#regression model selection
#stepwise (newer model) that adds a penalty for each useless variable
m2=step(m1)
summary(m2)
```

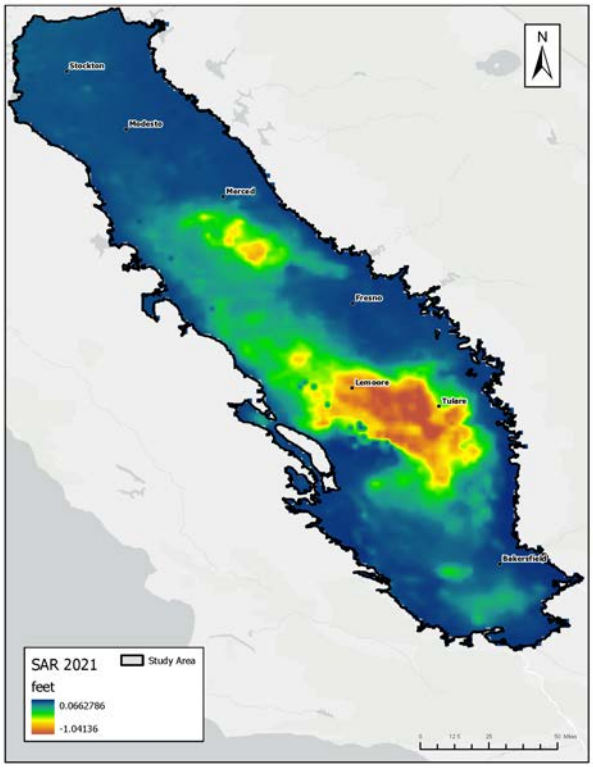
```
#removes one variable each time to see what variable has the least and largest impact
#we are actually applying BIC even though the output says AIC as the k adjusts (number of
variables) the values

n <- nrow(SJB)
m3 <- step(m1, k=log(n))
summary(m3)

#regression model diagnostics
par(mfrow=c(1,2))
plot(m3, which=c(1,2), cex=0.1)
```

Appendix E – Original SAR Land Subsidence Maps

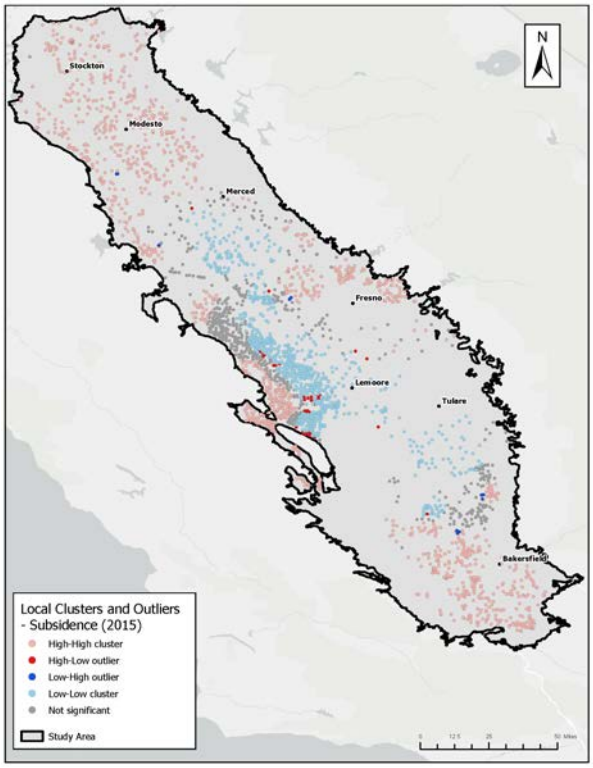


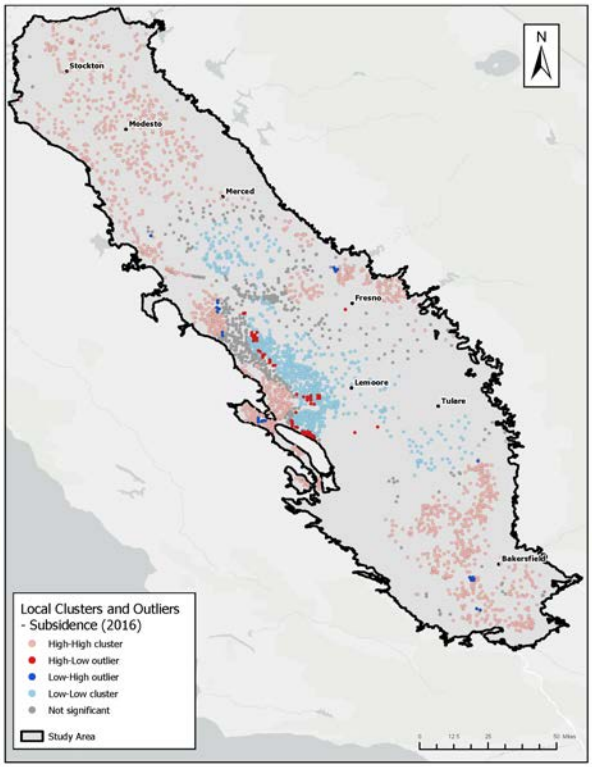

N
Stockton
Modesto
Merced
Fresno
Lemoore
Tulare
Bakersfield
SAR 2019
Study Area
feet
0.221832
-0.608073

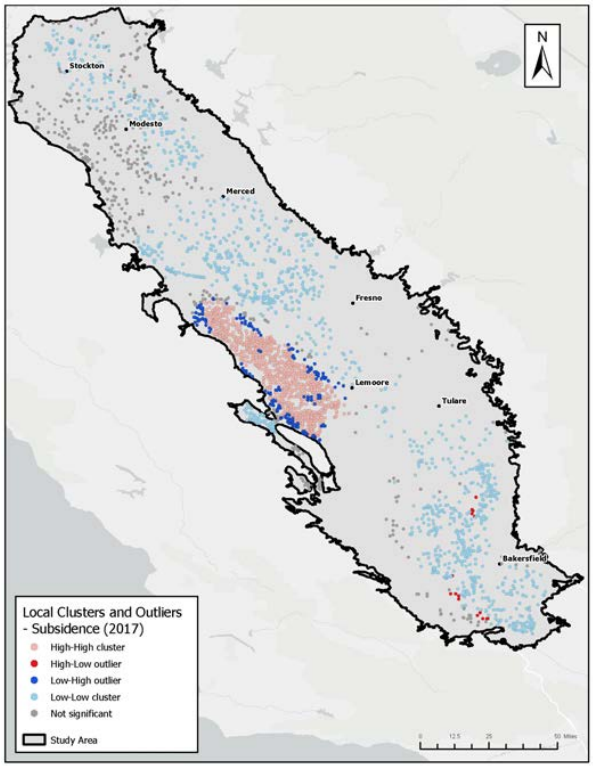

N
Stockton
Modesto
Merced
Fresno
Lemoore
Tulare
Bakersfield
SAR 2020
Study Area
feet
0.0689583
-0.87214

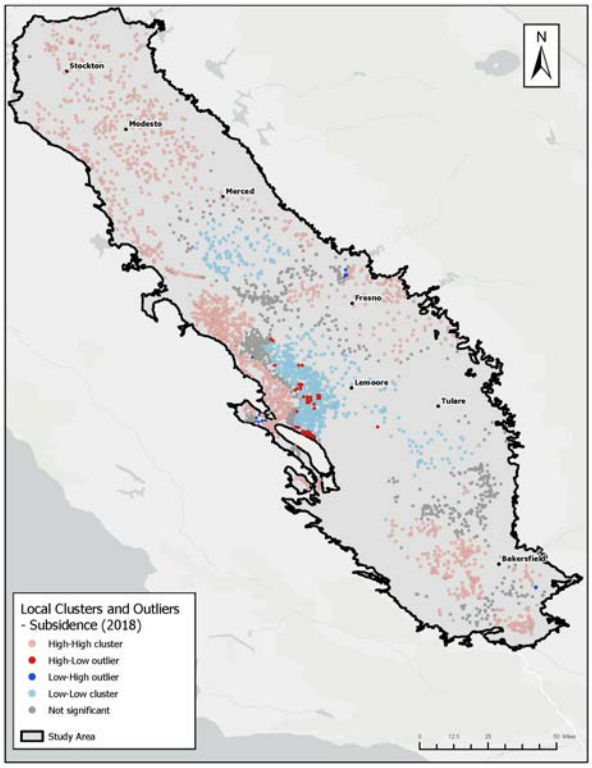

N
Stockton
Modesto
Merced
Fresno
Lemoore
Tulare
Bakersfield
SAR 2021
Study Area
feet
0.0662786
-1.04136

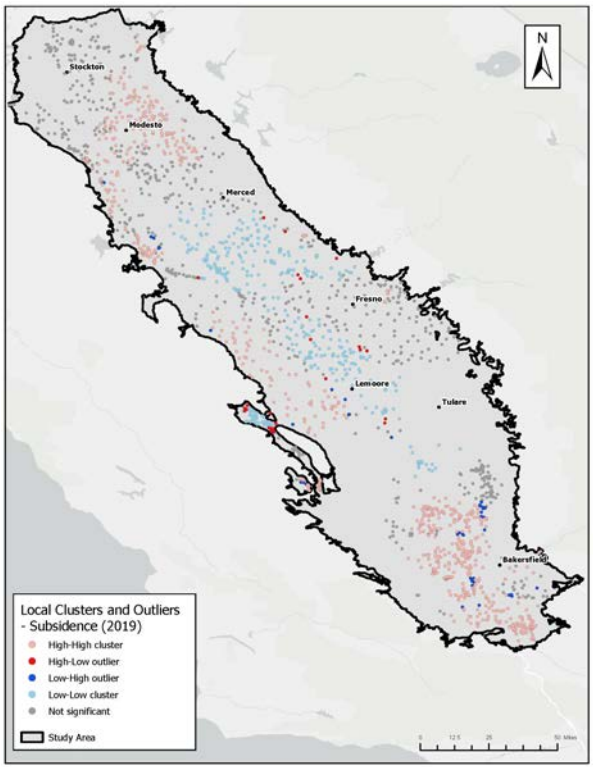
Appendix F – Moran's I Clusters and Outliers Maps

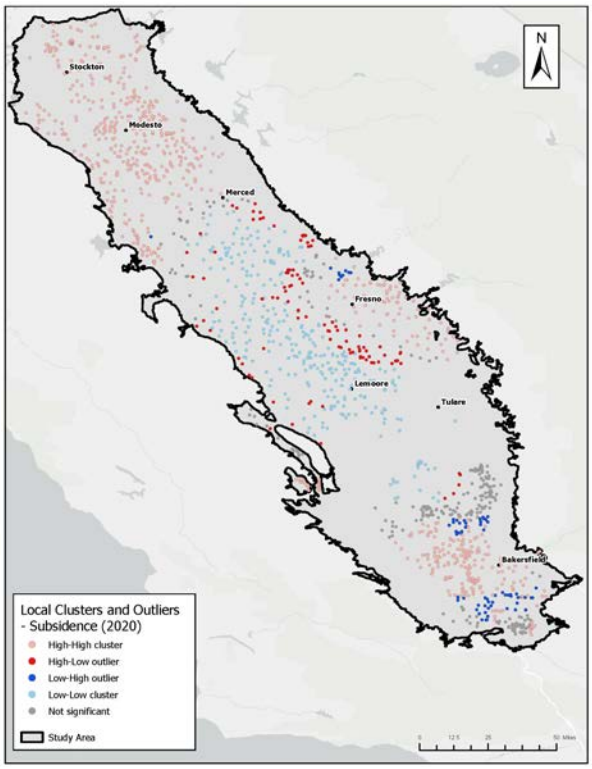


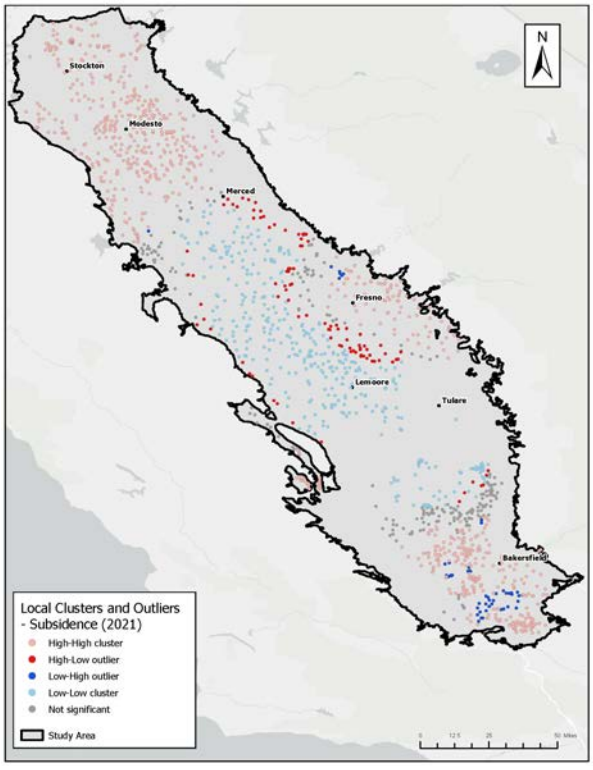




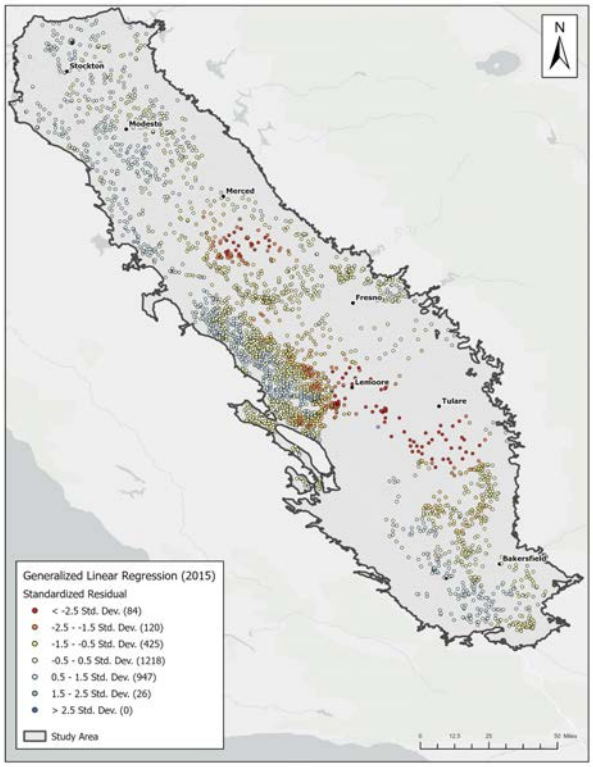


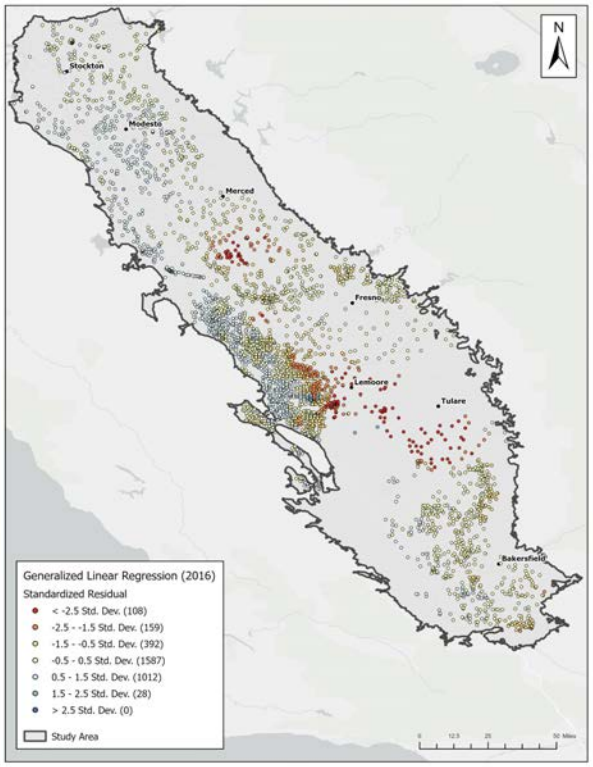

N
Stockton
Modesto
Merced
Fresno
Lemoore
Tulare
Bakersfield
Local Clusters and Outliers
- Subsidence (2019)
High-High cluster
High-Low outlier
Low-High outlier
Low-Low cluster
Not significant
Study Area

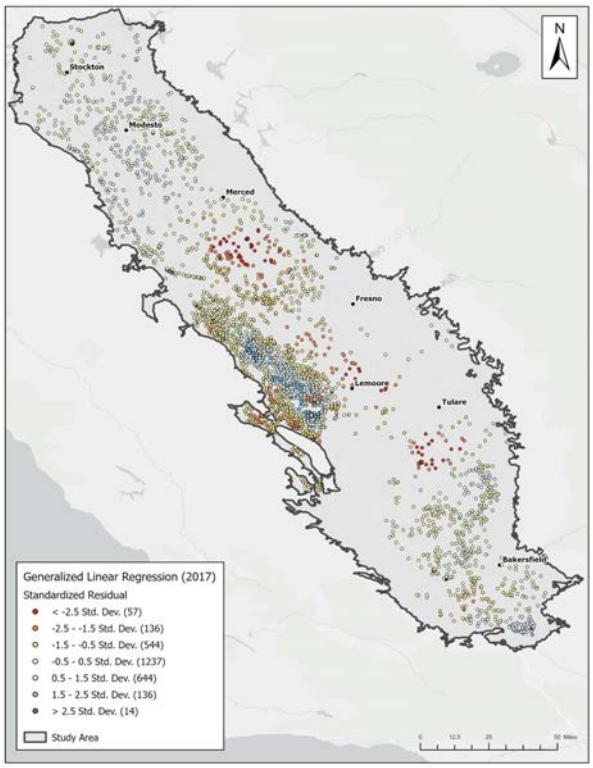

N
Stockton
Modesto
Merced
Fresno
Lemoore
Tulare
Bakersfield
Local Clusters and Outliers
- Subsidence (2020)
High-High cluster
High-Low outlier
Low-High outlier
Low-Low cluster
Not significant
Study Area

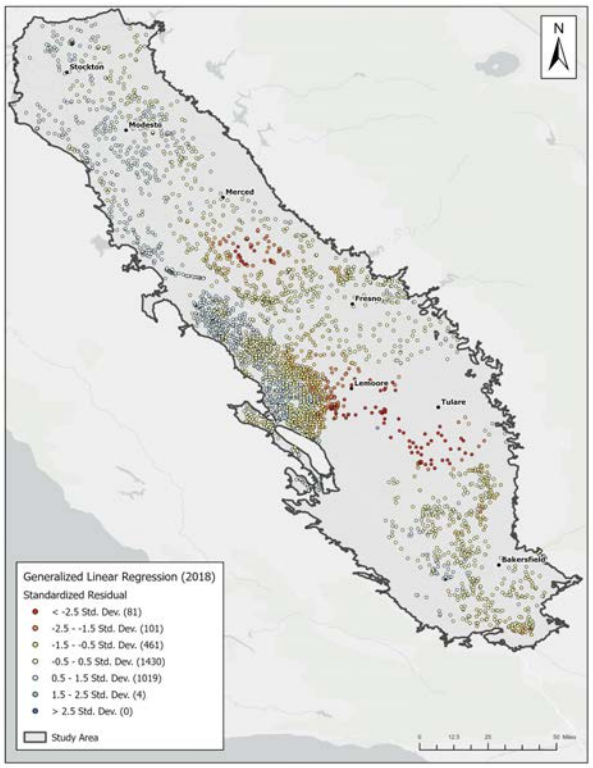

N
Stockton
Modesto
Merced
Fresno
Lemoore
Tulare
Bakersfield
Local Clusters and Outliers
- Subsidence (2021)
High-High cluster
High-Low outlier
Low-High outlier
Low-Low cluster
Not significant
Study Area

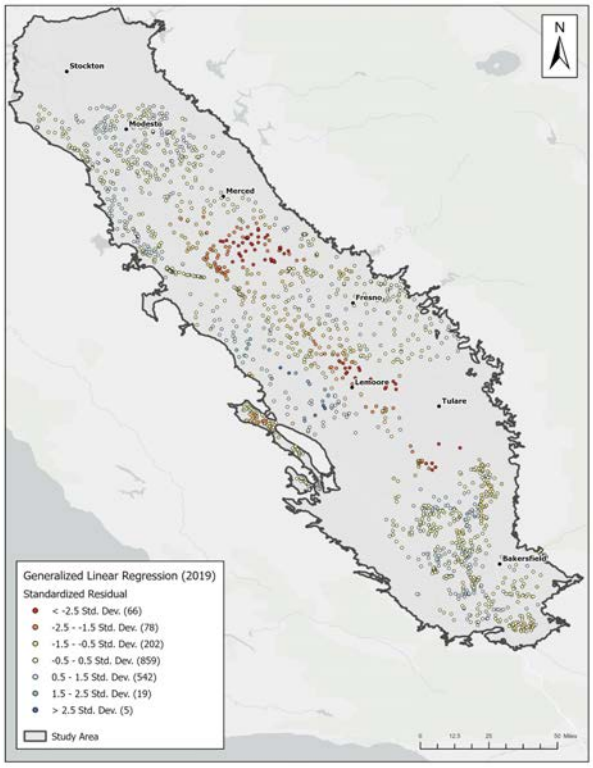
Appendix G – MLR Global Variable Coefficients Maps

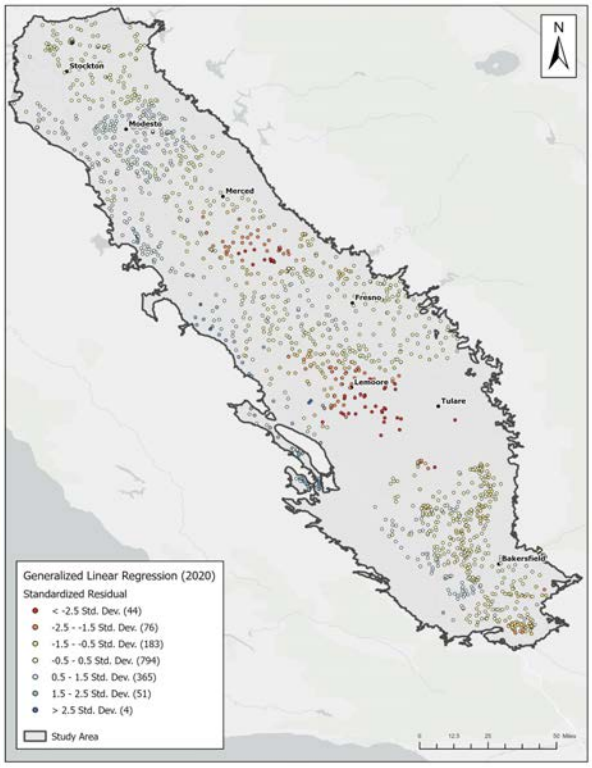


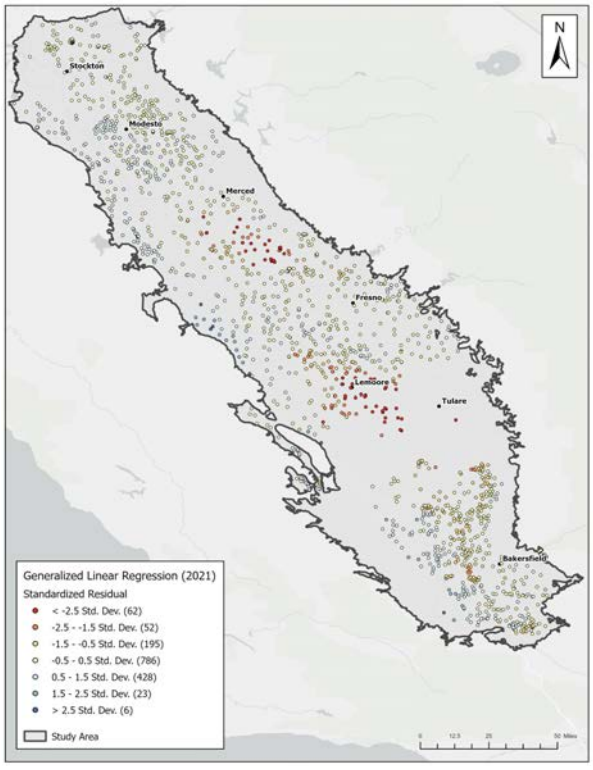




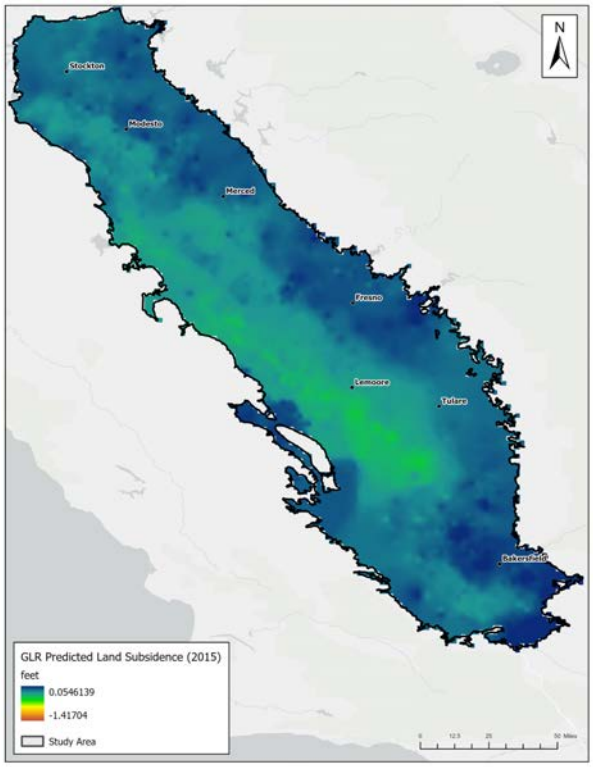


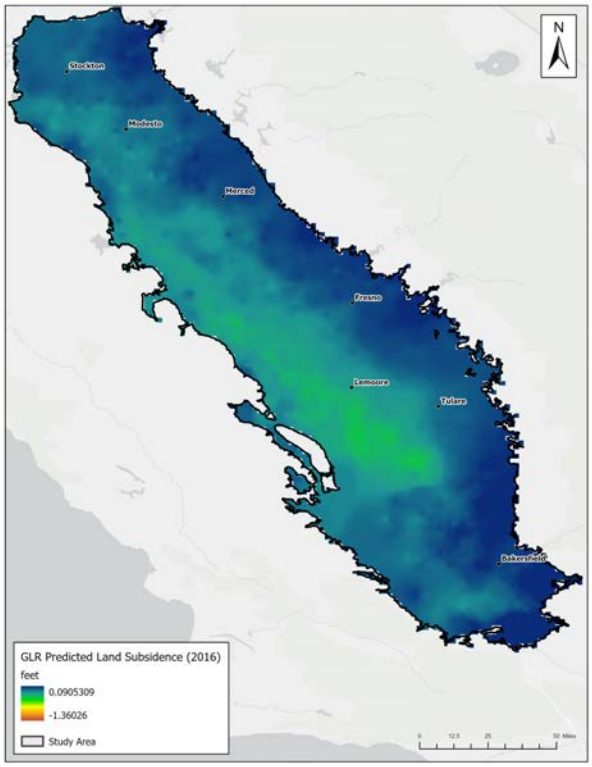

N
Stockton
Modesto
Merced
Fresno
Lemoore
Tulare
Bakersfield
Generalized Linear Regression (2019)
Standardized Residual
< -2.5 Std. Dev. (66)
-2.5 - -1.5 Std. Dev. (78)
-1.5 - -0.5 Std. Dev. (202)
-0.5 - 0.5 Std. Dev. (859)
0.5 - 1.5 Std. Dev. (542)
1.5 - 2.5 Std. Dev. (19)
> 2.5 Std. Dev. (5)
Study Area

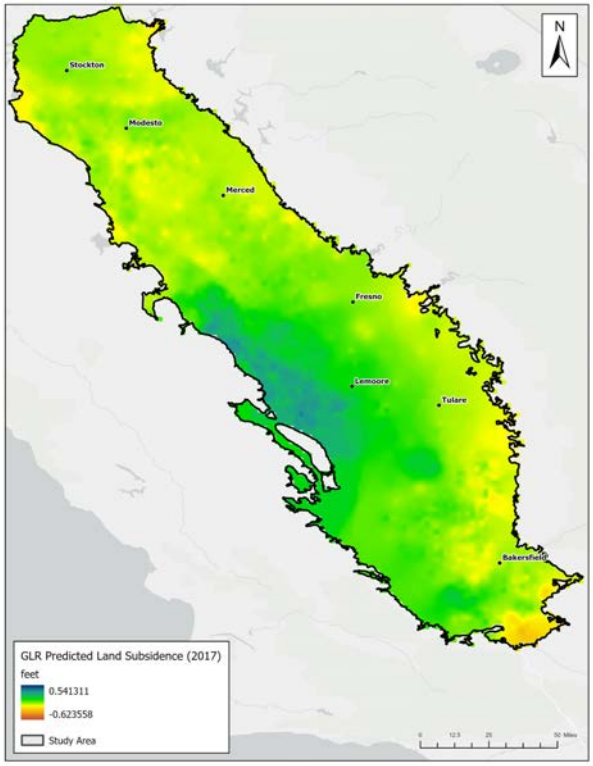

N
Stockton
Modesto
Merced
Fresno
Lemoore
Tulare
Bakersfield
Generalized Linear Regression (2020)
Standardized Residual
< -2.5 Std. Dev. (44)
-2.5 - -1.5 Std. Dev. (76)
-1.5 - -0.5 Std. Dev. (183)
-0.5 - 0.5 Std. Dev. (794)
0.5 - 1.5 Std. Dev. (365)
1.5 - 2.5 Std. Dev. (51)
> 2.5 Std. Dev. (4)
Study Area

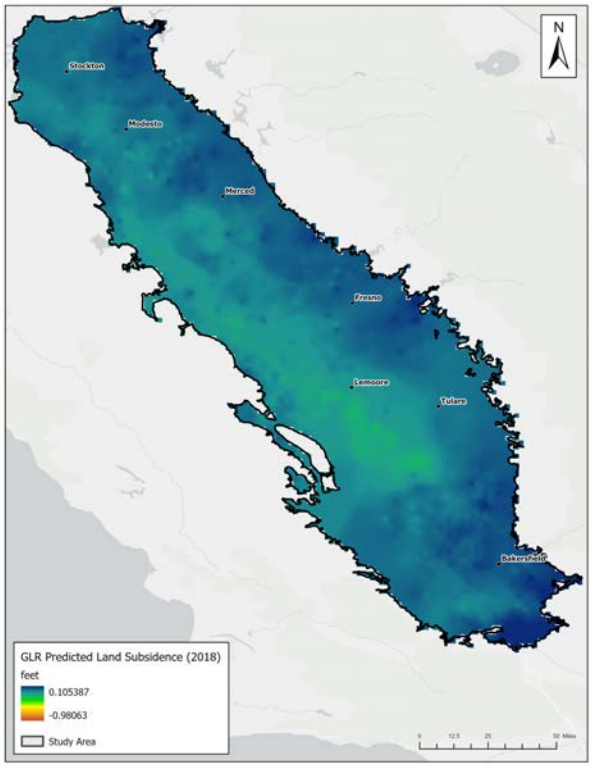

N
Stockton
Modesto
Merced
Fresno
Lemoore
Tulare
Bakersfield
Generalized Linear Regression (2021)
Standardized Residual
< -2.5 Std. Dev. (62)
-2.5 - -1.5 Std. Dev. (52)
-1.5 - -0.5 Std. Dev. (195)
-0.5 - 0.5 Std. Dev. (786)
0.5 - 1.5 Std. Dev. (428)
1.5 - 2.5 Std. Dev. (23)
> 2.5 Std. Dev. (6)
Study Area

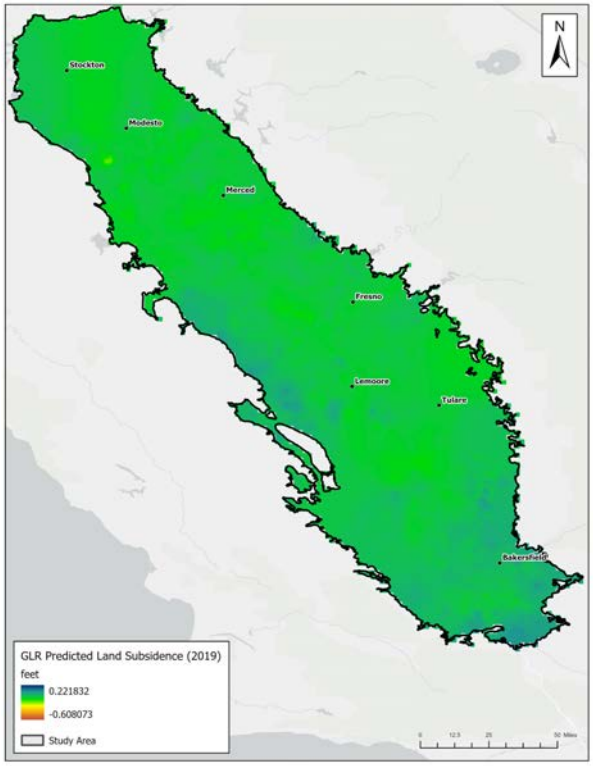
Appendix H– MLR Predicted Land Subsidence Maps

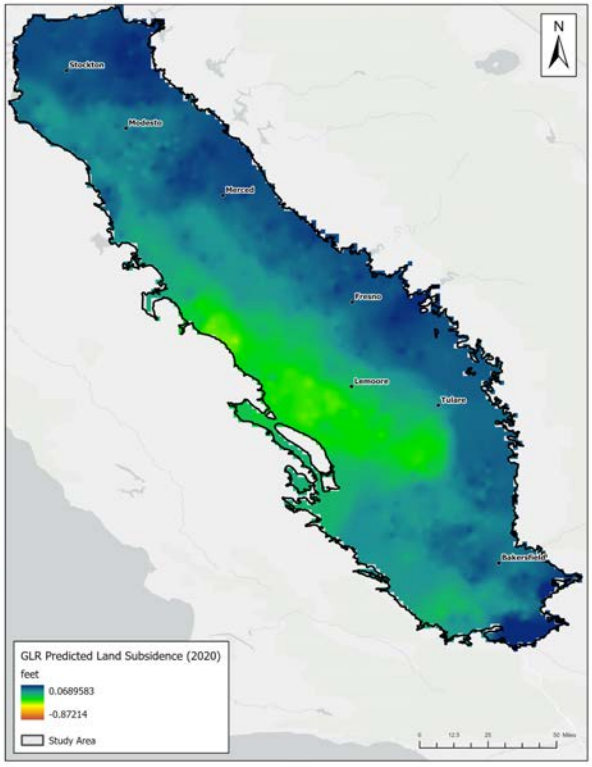


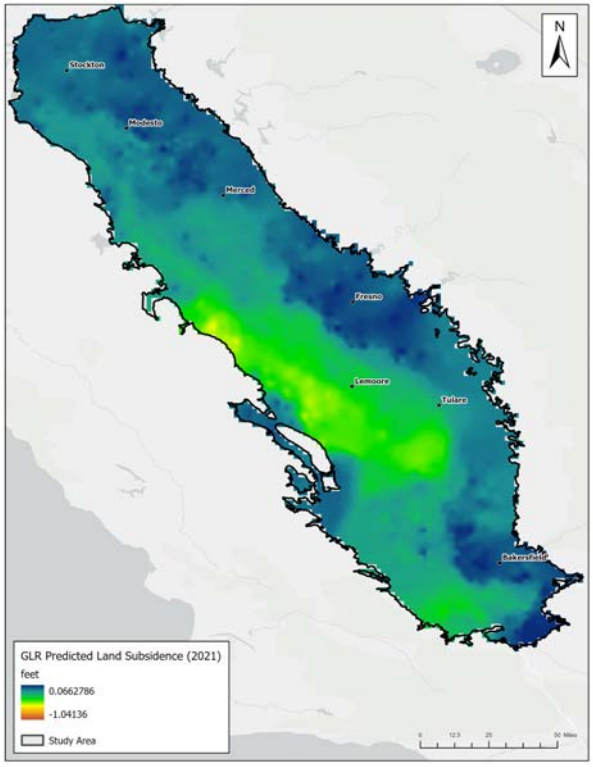




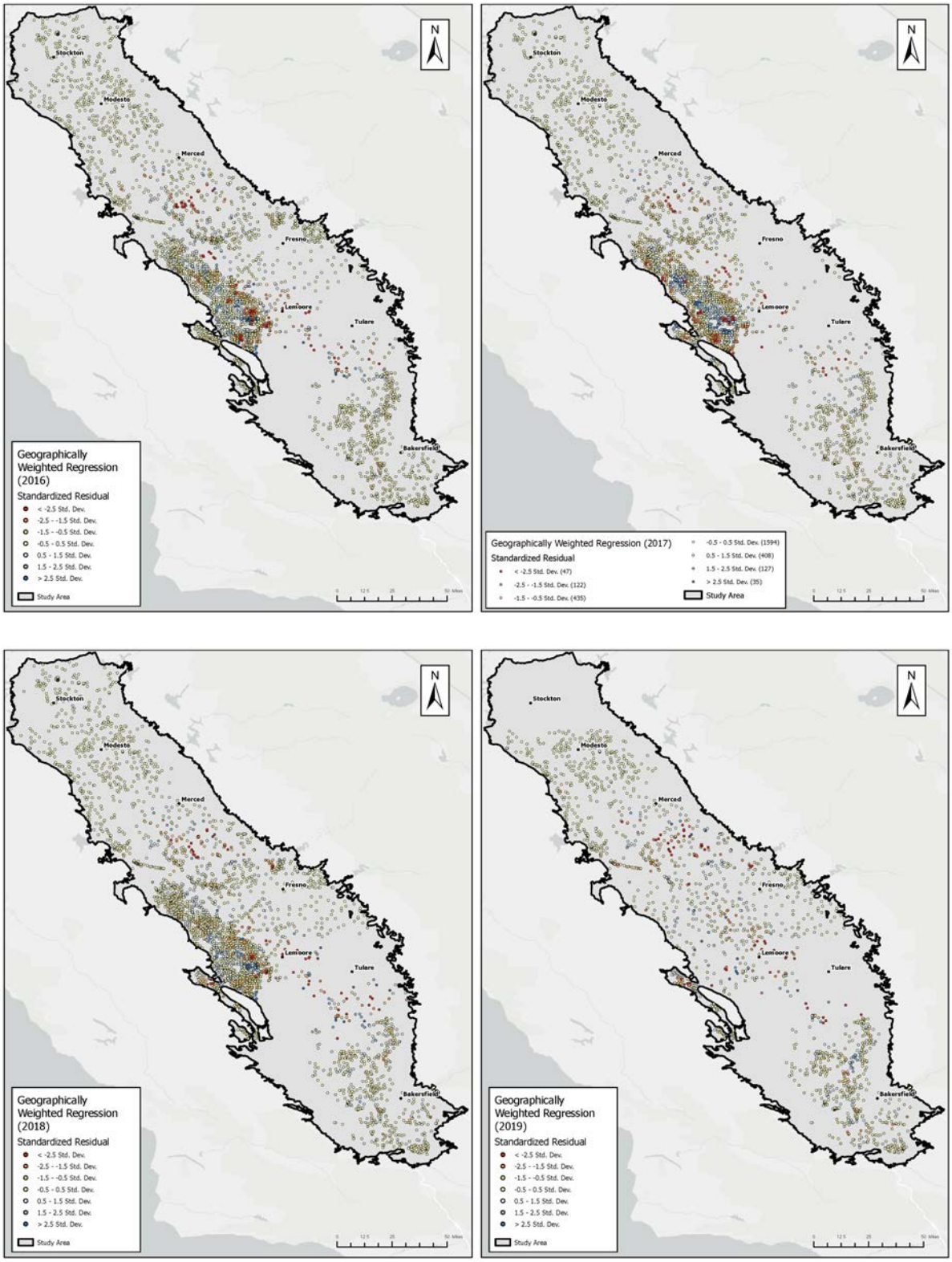


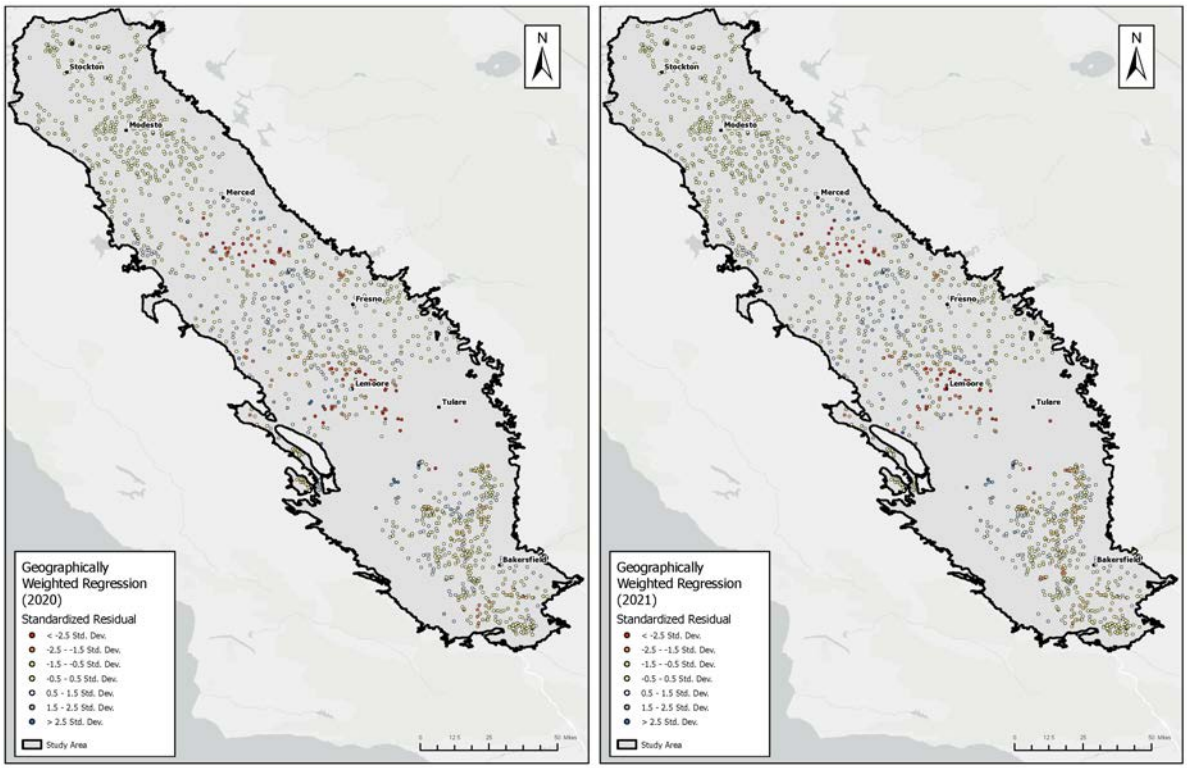

N
Stockton
Modesto
Merced
Fresno
Lemoore
Tulare
Bakersfield
GLR Predicted Land Subsidence (2019)
feet
0.221832
-0.608073
Study Area


N
Stockton
Modesto
Merced
Fresno
Lemoore
Tulare
Bakersfield
GLR Predicted Land Subsidence (2020)
feet
0.0689583
-0.87214
Study Area

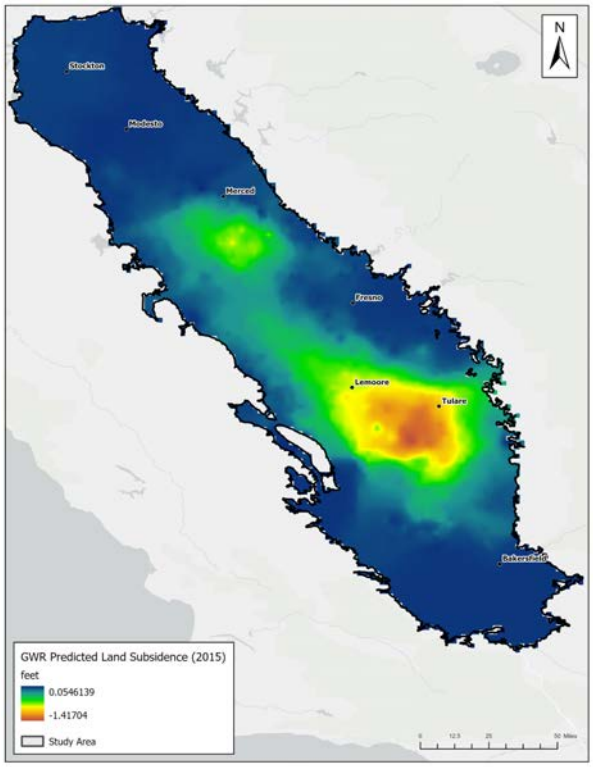

N
Stockton
Modesto
Merced
Fresno
Lemoore
Tulare
Bakersfield
GLR Predicted Land Subsidence (2021)
feet
0.0662786
-1.04136
Study Area

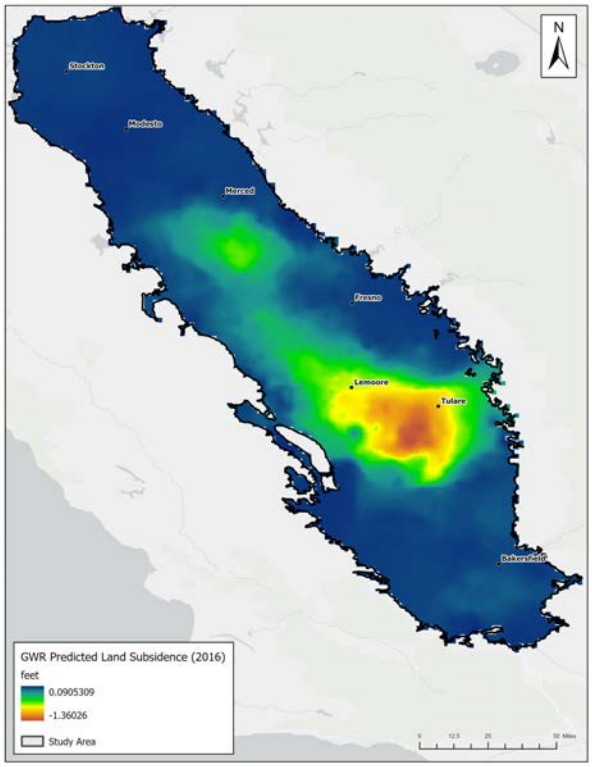
Appendix I – GWR Local Regression Residual Maps

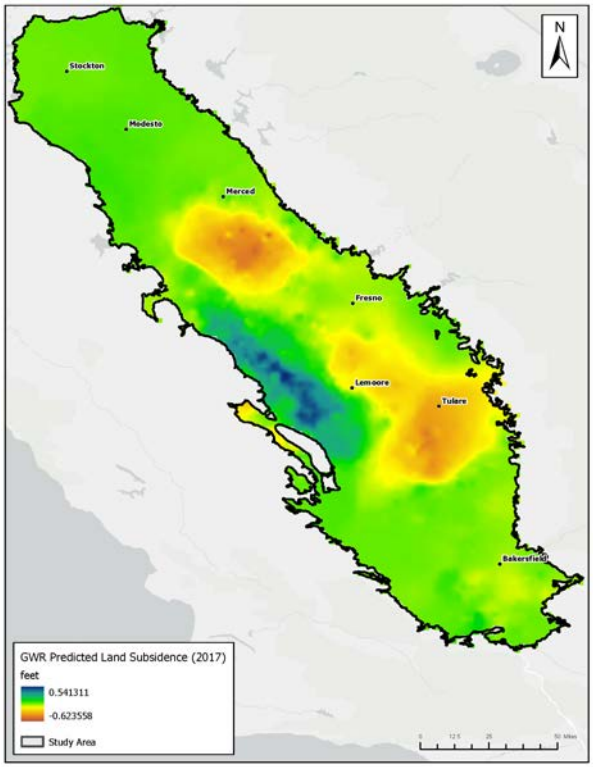


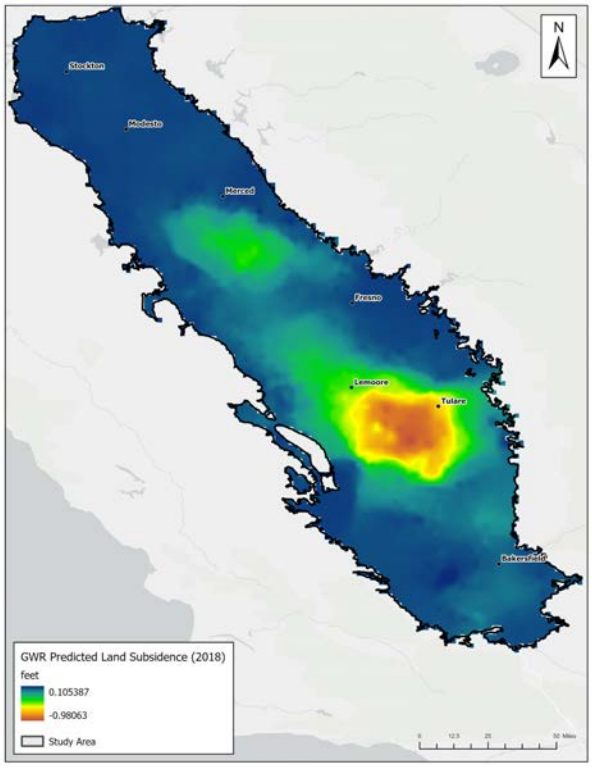

N
Stockton
Modesto
Merced
Fresno
Lemoore
Tulare
Bakersfield
Geographically Weighted Regression (2020)
Standardized Residual
< -2.5 Std. Dev.
-2.5 - -1.5 Std. Dev.
-1.5 - -0.5 Std. Dev.
-0.5 - 0.5 Std. Dev.
0.5 - 1.5 Std. Dev.
1.5 - 2.5 Std. Dev.
> 2.5 Std. Dev.
Study Area
0 12.5 25 50 Miles
N
Stockton
Modesto
Merced
Fresno
Lemoore
Tulare
Bakersfield
Geographically Weighted Regression (2021)
Standardized Residual
< -2.5 Std. Dev.
-2.5 - -1.5 Std. Dev.
-1.5 - -0.5 Std. Dev.
-0.5 - 0.5 Std. Dev.
0.5 - 1.5 Std. Dev.
1.5 - 2.5 Std. Dev.
> 2.5 Std. Dev.
Study Area
0 12.5 25 50 Miles

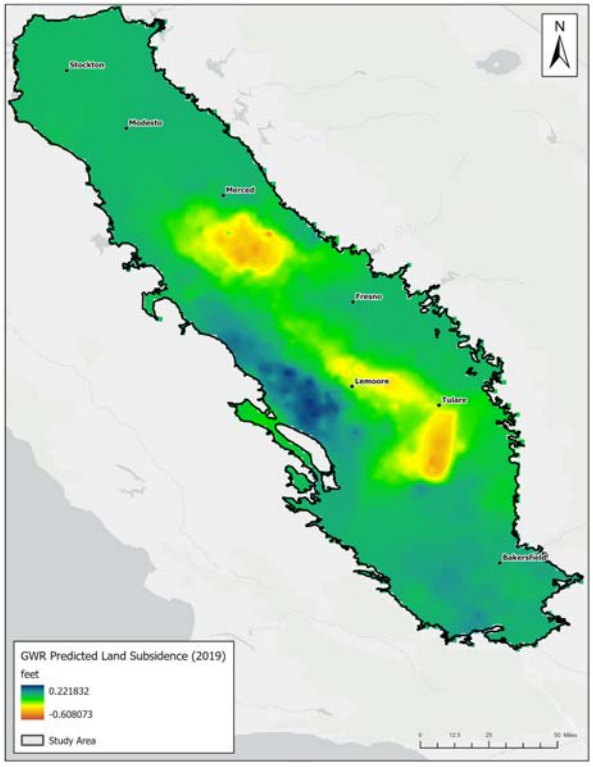
Appendix J – GWR Predicted Land Subsidence Maps

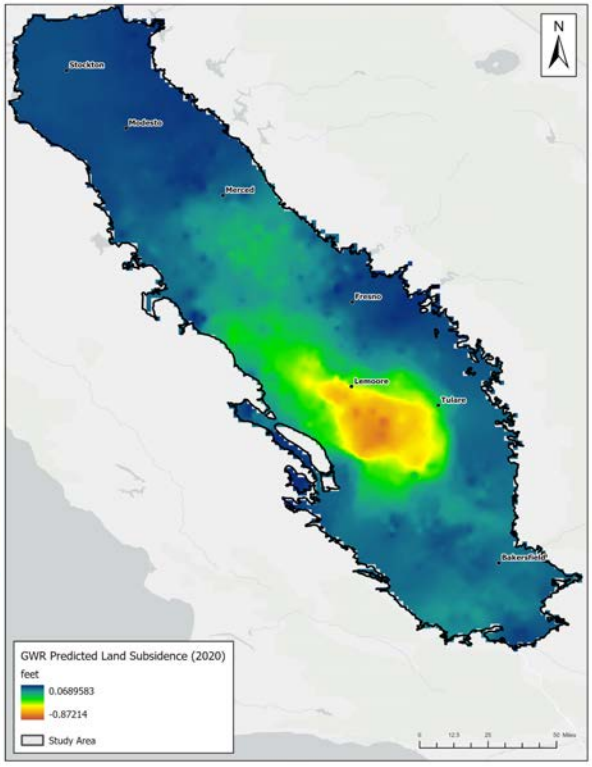


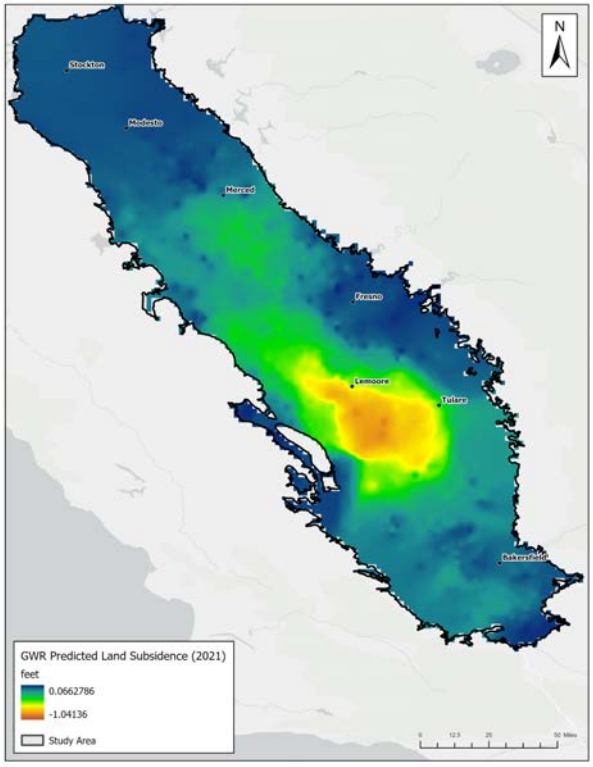







N
Stockton
Modesto
Merced
Fresno
Lemoore
Tulare
Bakersfield
GWR Predicted Land Subsidence (2019)
feet
0.221832
-0.608073
Study Area


N
Stockton
Modesto
Merced
Fresno
Lemoore
Tulare
Bakersfield
GWR Predicted Land Subsidence (2020)
feet
0.0689583
-0.87214
Study Area


N
Stockton
Modesto
Merced
Fresno
Lemoore
Tulare
Bakersfield
GWR Predicted Land Subsidence (2021)
feet
0.0662786
-1.04136
Study Area