California's Human Right to Water:
A Comparative Analysis of Domestic Well Water Quality Modeling

Claire A. Krumm

ABSTRACT

Several bills have been passed in California to fund community-level water quality improvement projects aimed at achieving the Human Right to Water. In an effort to produce a continuous model of water quality throughout the state, several research groups have utilized different modeling methods in terms of data sources used, time period from which water quality data was included, and geography over which individual water well measurements were aggregated. In this study, I used water quality data from GeoTracker GAMA to model concentrations of arsenic (As), chromium-6 (CR-6), 1,2,3-Trichloropropane (1,2,3-TCP), and nitrate as nitrogen (N). I controlled the parameters of time by comparing models using the past ten years of data to models using the past twenty years of data and geography by comparing models aggregating over sections and townships. Modeling choices impacted estimates of at-risk populations on the order of thousands to tens of thousands of people, although time affected these estimates to a lesser extent than geography. The two-pronged approach taken for 1,2,3-TCP demonstrated that the quality of existing data and the data cleaning choices made in the modeling process also largely impact final model estimates. Future water quality modeling efforts should incorporate a sensitivity analysis evaluating how different choices affect model outcomes and highlights the need for more complete and regular water quality testing throughout the state.

KEYWORDS

GeoTracker GAMA, Maximum Contaminant Level (MCL), likely Domestic Well Area (DWA), time-weighted average, at-risk population

INTRODUCTION

In 2012, California passed Assembly Bill (AB) 685 and became the first state to legally recognize the Human Right to Water. AB 685 states that all Californians have the right to "safe, clean, affordable, and accessible" water ("Human Right to Water" 2019). California impressively has one of the largest state-built, multi-purpose water systems in the United States, irrigating 750,000 acres of farmland and providing twenty-seven million of its citizens with drinking water ("Infrastructure" 2019). Despite its sophisticated water infrastructure, communities still suffer significant drinking water contamination. In 2017, roughly 50% of the 2,799 total violations recorded among public water systems (PWSs) were cited for failing to comply with Maximum Contaminant Levels (MCL) for regulated chemical contaminants or failing to follow treatment techniques outlined under the Safe Water Drinking Act (*2017 Annual Compliance Report* 2018). MCLs are public health regulations that determine minimum allowable concentrations of contaminants in drinking water ("Chemicals and Contaminants in Drinking Water | California State Water Resources Control Board" n.d.).

Resulting from the persistence of violations among public water systems, an estimated 6 million Californians have been served by systems that have been out of compliance at least once since 2012 ("Human Right to Water" 2019). Of additional concern are the estimated 1.5 – 2.5 million Californians who live outside of public water system boundaries and rely on water sources that draw on groundwater not subject to drinking water regulations, such as private domestic wells and small water systems with fewer than 5 service connections ("Human Right to Water" 2019). Importantly, complex patterns that arise in water quality violations across California that cannot simply be attributed to geography (Balazs et al. 2011, 2012). In many cases, *who* is served is a better indicator of water quality than *where* they are served.

Following the seminal 1987 report "Toxic Wastes and Race in the United States" revealing the discriminatory siting of toxic waste disposals, environmental justice research has revealed disparities in environmental quality persistent across the United States at the national, state, and local levels (Lee 1987, Cotton 2018, Carpenter and Wagner 2019, Meenar et al. 2019). Recent environmental justice research has confirmed that disparities in water quality are also correlated with social characteristics across the United States (McDonald and Jones 2018, Schaider et al. 2019) and in California specifically (Balazs et al. 2011, 2012). In terms of water justice, significant

social characteristics include the following: (1) socioeconomic status (McDonald and Jones 2018, Wikstrom et al. 2019, Schaider et al. 2019); (2) percent home ownership (Balazs et al. 2011, 2012, McDonald and Jones 2018, Schaider et al. 2019); and (3) percent minority (Balazs et al. 2011, 2012, McDonald and Jones 2018, Schaider et al. 2019). A consistent finding between water justice studies was that correlative results will vary depending on the scale of study, suggesting scale matters when analyzing water quality data (Patrick et al. 2014, Schaider et al. 2019). Given these patterns, it is vital to analyze California's water quality data in a way that produces equitable solutions for water infrastructure that is discriminatorily failing across the state.

In recognition of the state's water infrastructure challenges, Senate Bill (SB) 200 established the Safe and Affordable Water Fund to allocate $130 million of funding for water infrastructure development to communities served by unsafe drinking water (Monning et al. 2019). SB 200 requires the State Water Resource Control Board to develop a fund expenditure plan that prioritizes disadvantaged communities served by a public water system and low-income households served by domestic wells (Monning et al. 2019). To evaluate how to best prioritize proposed projects, a complete understanding of the drinking water quality landscape across California is required, particularly with regard to areas served by domestic wells. Currently, the best available statewide data on the quality of groundwater is housed in GeoTracker GAMA, which is a repository for groundwater samples collected by California's State Water Resources Control Board, California's Department of Water Resources, and various other water monitoring groups ("GAMA Groundwater" 2019). GeoTracker GAMA is a valuable source of water quality data, yet there are still significant gaps in water sampling data. These gaps include a paucity of water quality measurements at different aquifer depths and water quality measurements that lack accompanying depth data. Additionally, the data is inconsistent in its geographic coverage for different contaminants. Because of the incomplete nature of the GAMA dataset, some modeling is required to interpolate statewide estimates of water quality for residents reliant on unregulated water sources.

There are several water quality models completed or currently underway by different research groups, including projects being undertaken by Sacramento State's Office of Water Programs ("California Groundwater Risk Index (GRID)" 2018), Cal EPA's Office of Environmental Health Hazard Assessment ("CalEnviroScreen 3.0" 2016), UC Berkeley's Water Equity Science Shop (WESS, "The Drinking Water Tool, 2019" 2019), and the California State

Water Resources Control Board (*Methodology to Estimate Groundwater Quality Accessed by Domestic Wells in California* 2019). Although these efforts have all relied on data for source water available through GeoTracker GAMA, each research group has made a different combination of decisions regarding (1) the range of years of GAMA data to include in the model, (2) the range of well depths from which data should be included, and (3) the geographical resolution over which samples should be averaged or aggregated. These decisions likely have significant impacts on how the model behaves in any given region in terms of the estimated contaminant concentrations and whether or not the area exceeds MCL standards. By extension, these decisions will also impact estimates of the number of people in California who are exposed to contaminant concentrations from groundwater at levels that exceed the MCL, which could in turn impact the allocation of state funding to address issues of water remediation and state efforts to achieve the goal of universal access to clean water.

The purpose of this research was to evaluate how modelling choices impact the estimates of people at risk of exposure to water exceeding the MCL. For this study, the following four contaminants were used: Arsenic (As), Chromium-6 (CR-6), 1,2,3-Trichloropropane (1,2,3-TCP), and Nitrate as Nitrogen (Nitrate as N). To evaluate these effects, I manipulated the parameters of time and geography, producing four models for comparison for each of the four contaminants. The models are evaluated on their estimations of (1) the location of areas exceeding the MCL, (2) total area exceeding the MCL, and (3) number of people at-risk of relying on source water that exceeds the MCL in the absence of water treatment.

METHODS

Study system

This study explores the effect of model choices on estimates of population risk. I developed statewide models to estimate the groundwater concentration of As, CR-6, 1,2,3-TCP, and N. The model represents source water, such as the water relied on private domestic wells in the absence of water treatment. In developing this model, I manipulated the parameters of (1) *time scale* by controlling the range of years from which data was included, and (2) *geography* by controlling the spatial area over which samples were averaged. The models were then applied to an existing data

layer estimating the population served by domestic wells in California ("The Drinking Water Tool, 2019" 2019). Using these two data sets (i.e. water quality and population served by domestic wells), I determined the population reliant on groundwater estimated to contain contaminant concentrations exceeding their corresponding Maximum Contaminant Level (MCL) for each contaminant under each set of model parameters.

This study incorporated public data from domestic and supply wells from GeoTracker GAMA, an online repository for groundwater quality data curated by the California Water Boards ("GAMA Groundwater" 2019). The GeoTracker GAMA data is provided by different authorities that range from public authorities, such as the Department of Water Resources, to private parties, such as community-lead local groundwater projects ("GAMA Groundwater" 2019). I included the GeoTracker GAMA datasets that sampled domestic wells or supply wells representative of domestic well water (Appendix A).

Modeling Rationale

The earliest measurements available in the GeoTracker GAMA datasets varied by contaminant and range from 1907 (N) to 1994 (CR-6). From year to year, samples are collected from wells located in different parts of the state. Thus, including samples from a larger range of years results in more data points and more geographic coverage. Using a multi-year range also accounts for the variation observed in contaminant concentrations over time, which is influenced by human activity and natural phenomena. For instance, the main source of N contamination in watersheds is fertilizer runoff ("Nitrates and Nitrites in Drinking Water | California State Water Quality Control Board" 2019). Therefore, N concentration is affected by the level of agricultural activity local to the well measured, which can change over time.

Although including data from more years increases the number of data points and accounts for temporality, the quality of GeoTracker GAMA data has varied throughout time due to technological changes in detection techniques. Water quality instruments have concentration detection limits, below which a contaminant's concentration cannot be accurately measured. As this technology has improved, the measurements have become more accurate, especially at lower contaminant concentrations. Data from earlier reporting years were often recorded with a qualifier,

such as a less than sign ('<'), indicating that the concentration is less than a minimum value but a more accurate value could not be measured due to instrument or assay limitations.

Water quality policy dictates MCLs at the local, state, and federal levels. MCLs are determined based on health and safety, achievability, and economic impact. As public health research advances, the contaminant concentration deemed "safe" continues to evolve. Additionally, technological advances change the reliability of testing, availability of water treatment options, and associated cost of achieving safe water. Therefore, in our dataset, samples collected over a twenty-year time span may be recorded with different reporting limits.

In addition to technological ability and standards, the water wells sampled in a given year do not cover all of California. Existing water wells are not uniformly distributed across California's geography, and are instead concentrated in the more populated and developed areas of the state. Beyond the underlying distribution of water wells available for sampling, not all water wells are sampled every year. Due to these inconsistencies in the water quality data available for California, modeling is required to interpolate statewide estimates.

There are two modeling choices that will likely increase geographic coverage: (1) including older data (using a larger date-range), and (2) aggregating water well samples over larger units of geography. To test this hypothesis and the effects these choices have on modeling outcomes, this study compares models using the past ten years of data to models using the past twenty years of data. It also compares the effects of aggregating over townships to aggregating over sections. Townships and sections are a grid system developed by the Bureau of Land Management ("Township and Range Survey System" n.d.). The typical township is a six mile by six mile area, while the typical section is one mile by one mile area ("Township and Range Survey System" n.d.).

Modeling methods

I downloaded the domestic and supply well datasets for each contaminant from GeoTracker GAMA on November 13, 2019 (Step 1, Figure 1). Using R (R Core Team 2019), I filtered the datasets by date to include the past ten years of data (2010-2019) and the past twenty years of data (2000-2019) (Step 2, Figure 1). I cleaned the data using the contaminant concentration, qualifier, and reporting limit for each measurement in the ten-year and twenty-year datasets (Appendix B).

Next, I grouped the data by unique well ID and year and averaged the contaminant concentration values (Step 3, Figure 1). This produced one contaminant concentration value per well ID per year. Finally, I grouped the data by unique well ID and averaged the contaminant concentration values across all years. This produced a single, time-weighted contaminant concentration value for each unique well ID in the ten-year and twenty-year datasets for each contaminant.

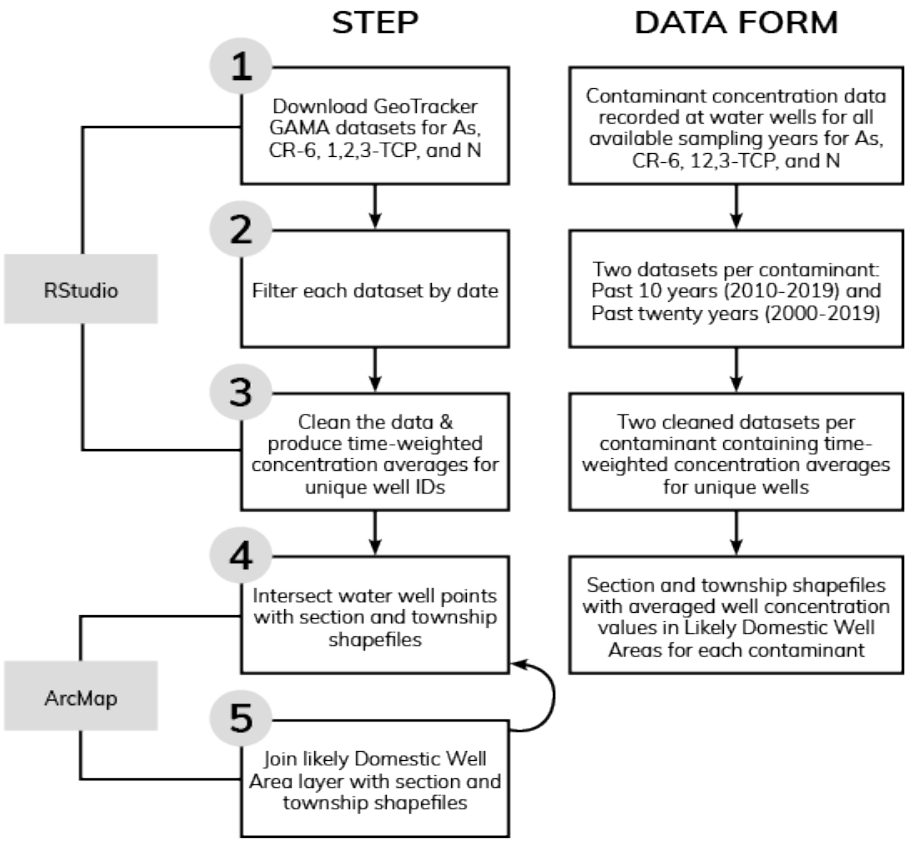

Figure 1: Conceptual diagram of methods process.

Using ArcMap (v. 10.8 ESRI 2020), I plotted the location of each well based on its latitude and longitude in decimal degrees. I intersected the well points with the township and section shapefiles provided by the Bureau of Land Management ("Cadastral/BLM_Natl_PLSS_CadNSDI (MapServer)" n.d.). I averaged the contaminant concentration values for wells located in the same township or section (Step 4, Figure 1). This produced four models, which I refer to with the following notation:

- TS10: The past ten years of data aggregated over townships.
- TS20: The past twenty years of data aggregated over townships.
- SECT10: The past ten years of data aggregated over sections.
- SECT20: The past twenty years of data aggregated over sections.

I then assigned population data to each model (TS10, TS20, SECT10, SECT20) using the population layer developed by WESS (Pace et al. 2019) which estimates populations reliant on domestic water wells in California at the section level based on four underlying datasets: US Census data ("TIGER/Line Shapefiles" n.d.), boundaries from the Tracking California Water System Service Areas tool ("Water Systems --- Tracking California" n.d.), well locations from the Online System for Well Completion Reports dataset ("Well Completion Reports" n.d.), and LandVision dataset locating parcels in California ("Real Estate Analysis & Mapping Application | LandVision™" n.d.).

For the section models, I joined my section layers containing water quality to the Domestic Well Area (DWA) layer by section code (MTRS). For the township models, I dissolved the DWA layer by township code (MTR), which is encoded in the first seven characters of the section code. I then joined the DWA layer to the township shapefile by the MTR field and summed the likely DWA populations within each township. I joined the township layers to the dissolved DWA layer via township codes. I developed the distribution maps and the percent change maps using ArcMap. Finally, I exported the population-assigned data tables from ArcMap to RStudio to obtain the descriptive statistics.

To produce the maps displaying percent change in water quality, I calculated the percent change in average concentration for a township comparing the twenty-year models to the ten-year models. To illustrate this method, consider Township M07S21E and the different percent changes the sections contained within it exhibit (Figure 2). The sections displayed with a prominent black border (1, 2, 4-10) contain both (1) at least one well with data from the past ten years dataset and (2) a likely DWA. For sections 1, 2, 4, 7, and 8, limiting the dataset to more recent data (from SECT20 to SECT10) resulted in an increase in the average N concentration of up to 50%. For sections 6 and 9, no SECT20 wells were filtered out when using the SECT10 model, and the wells did not have any older concentration measurements (2000-2010) that altered the time-weighted average for sections 6 and 9. Therefore, these sections exhibited no change in average concentration.

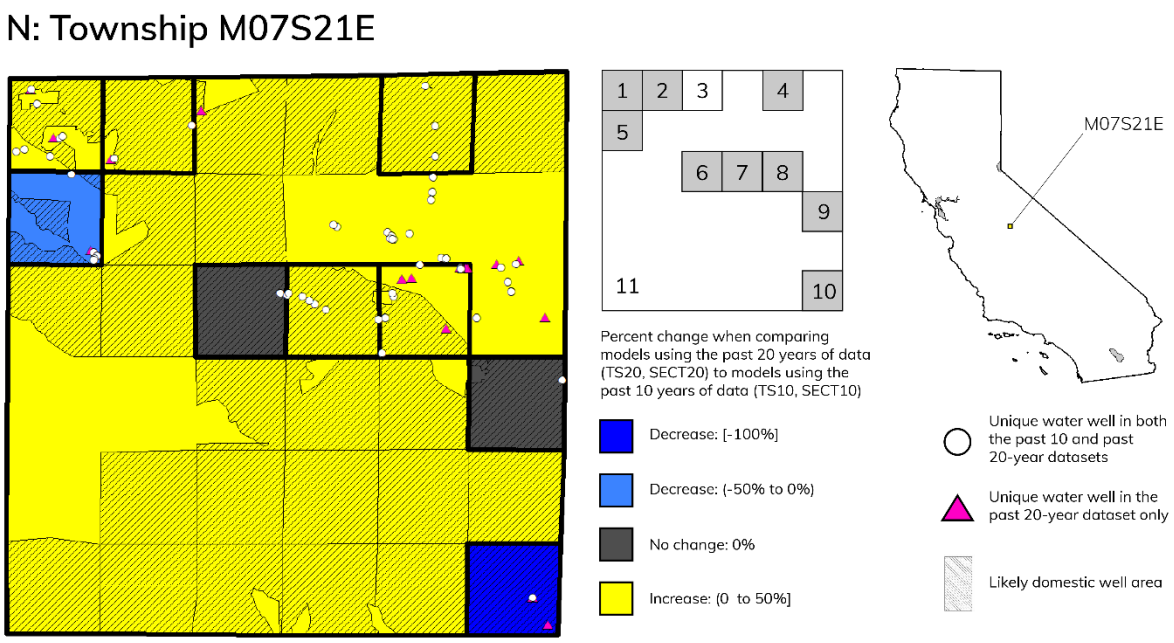

Figure 2: Example displaying the percent changes in N concentrations for a township (MTR: M07S21E) and the sections it encloses. The sections with percent change values are boldly outlined in the colored diagram and filled in with gray in the schematic to the right (sections 1, 2, 4-10). Section 3 does not have a percent change value, although it does have a water well with concentration data in the past twenty-years dataset. Those sections for which moving from a dataset including older sampling data (2000-2019) to a dataset including only more recent sampling data (2010-2019) were excluded from the percent change maps and are instead represented in the data availability maps.

In section 3, there was a well in the SECT20 dataset that was not present in the SECT10 dataset. These cases, in which a section or township went from containing a well with concentration data to not containing a well when using the more recent dataset, have been removed from the percent change maps. Instead, they are displayed in the data availability maps, which highlight sections and townships that gained data when using a dataset including older samples (Figures 3c, 3d, 4c, 4d, 5c, 5d, 6c, 6d, 7c, 7d). Although sections 5 and 10 experienced percent decreases and sections 6 and 9 experienced no change, the entire township experienced an increase in N concentration when comparing the twenty-years dataset to the ten-years dataset. This example demonstrates how changes at the section level are not always captured at the township level.

The DWA layer was developed to estimate domestic well water reliance at the section level. When aggregating to a larger spatial resolution, some townships will contain likely DWAs and water wells with concentration data that lie outside of the likely DWA boundaries. Because the unit of observation for the TS10 and TS20 models is the township, all wells located within the township were used to calculate the time-weighted contaminant concentration averages. The

population for each township was calculated by summing the populations of all of the sections containing likely DWAs within the township, even if there was no concentration data from GeoTracker GAMA for that particular DWA.

RESULTS

For all contaminants, the median contaminant concentration for each model was below the MCL (Table 1): As, 10 µg/L; CR-6, 10 µg/L; 1,2,3-TCP, 0.005 µg/L; N, 10 mg/L. N, the contaminant with the greatest number of water wells with data, also had the most geographic coverage in terms of number of sections (SECT10: 3,936, SECT20: 4,634) and the number of townships (TS10: 1,513, TS20: 1,607) with data in likely DWAs. The second method for 1,2,3-TCP (1,2,3-TCP_2), in which concentration values lacking a recorded reporting limit were removed, resulted in the least number of townships and sections with concentration data in likely DWAs (SECT10: 238, SECT20: 432, TS10: 225, TS20: 445). These smaller inputted datasets for 1,2,3-TCP_2 resulted in larger percentages of townships and sections with concentrations at or exceeding the MCL compared to 1,2,3-TCP_1 (in which values lacking reporting limits were not removed), but resulted in fewer estimated at-risk people than with the models for 1,2,3-TCP_1.

Overall, the percentage of area with concentrations at or exceeding the MCL is similar to that of the percentage of townships and sections at or exceeding the MCL. This is due to the consistency in section and township area, where sections are typically 1 mi^2 (1 mi x 1 mi) and townships are typically 36 mi^2 (6 mi x 6 mi). However, when comparing models for a given contaminant, the absolute estimates of at-risk populations was largely variable depending on the modeling approach.

Table 1: Descriptive statistics of average contaminant concentration for townships and sections located in Likely Domestic Well Areas. Data was downloaded from the California Waterboard's GeoTracker GAMA database. The MCLs for the contaminants are as follows: As, 10 µg/L; CR-6, 10 µg/L; 1,2,3-TCP, 0.005 µg/L; N, 10 mg/L.

Contaminant (MCL)	Model	Median	IQR	95th percentile	Number townships or sections with concentration data	Percentage of townships or sections with concentrations ≥ MCL	Population in areas with concentrations ≥ MCL	Total area (km^2) with concentration data	Percentage of area with concentrations ≥ MCL
As (10 µg/L)	SECT10	1.33	4.2	18.4	2,802	9.96%	54,358	7,287	9.95%
	SECT20	1.34	3.93	17.6	3,574	9.82%	69,688	9,306	9.85%
	TS10	1.46	4.03	19.6	1,350	11.11%	113,354	123,312	11.18%
	TS20	1.59	3.95	19.8	1,482	11.13%	127,077	135,238	11.32%
CR-6 (10 µg/L)	SECT10	0.103	2.1	8.09	1,957	4.04%	18,789	5,086	4.03%
	SECT20	0.17	2.1	8.53	2,057	4.04%	18,543	5,347	4.02%
	TS10	0.3	1.74	7.88	1,101	3.36%	28,043	100,548	3.35%
	TS20	0.33	1.75	8.04	1,129	3.45%	27,783	103,162	3.44%
1,2,3-TCP_1 (0.005 µg/L)	SECT10	0	0	0.0121	2,233	6.90%	28,512	5,797	6.93%
	SECT20	0	0	0.00658	2,862	5.56%	33,824	7,434	5.58%
	TS10	0	0	0.00705	1,193	5.87%	158,918	108,921	6.01%
	TS20	0	0	0.00616	1,347	5.27%	137,201	122,902	5.42%
1,2,3-TCP_2 (0.005 µg/L)	SECT10	0	0	0.0747	238	11.34%	1,961	622	11.41%
	SECT20	0	0	0.045	432	9.49%	4,052	1,131	9.55%
	TS10	0	0	0.0675	225	16.00%	71,886	20,858	16.16%
	TS20	0	0	0.0406	445	11.01%	75,525	41,253	11.12%
N (10 mg/L)	SECT10	1.5	3.93	14.4	3,936	8.33%	46,790	10,238	8.35%
	SECT20	1.48	3.96	14.2	4,634	8.37%	48,203	12,061	8.37%
	TS10	1.13	3.05	8.87	1,513	3.90%	57,514	137,765	3.87%
	TS20	1.11	2.86	8.75	1,607	4.04%	69,398	146,071	4.07%

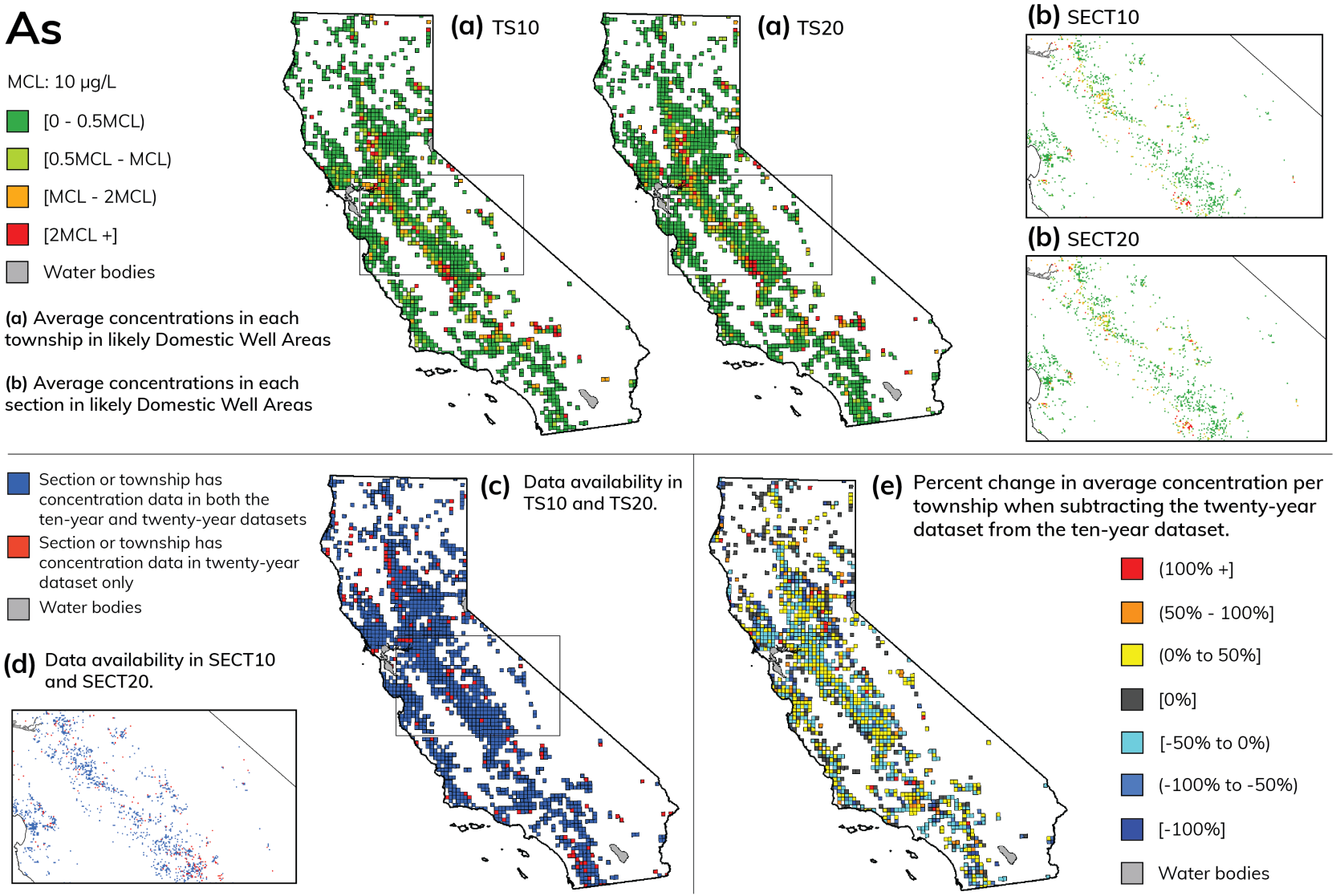

Figure 3: Maps displaying the results of the As models. 3a: Average concentration in each township containing a likely DWA. 3b: Average concentration in each section containing a likely DWA. 3c: Depiction of townships that have concentration data in the TS10 and TS20 models. 3d: Depiction of sections that have concentration data in the SECT10 and SECT20 models. 3e: Percent change in township concentration when comparing TS20 to TS10 estimates. E.g. An increase of 20% (yellow) indicates that when removing 2000-2010 sampling data for the ten-years dataset, the average concentration increased by 20%.

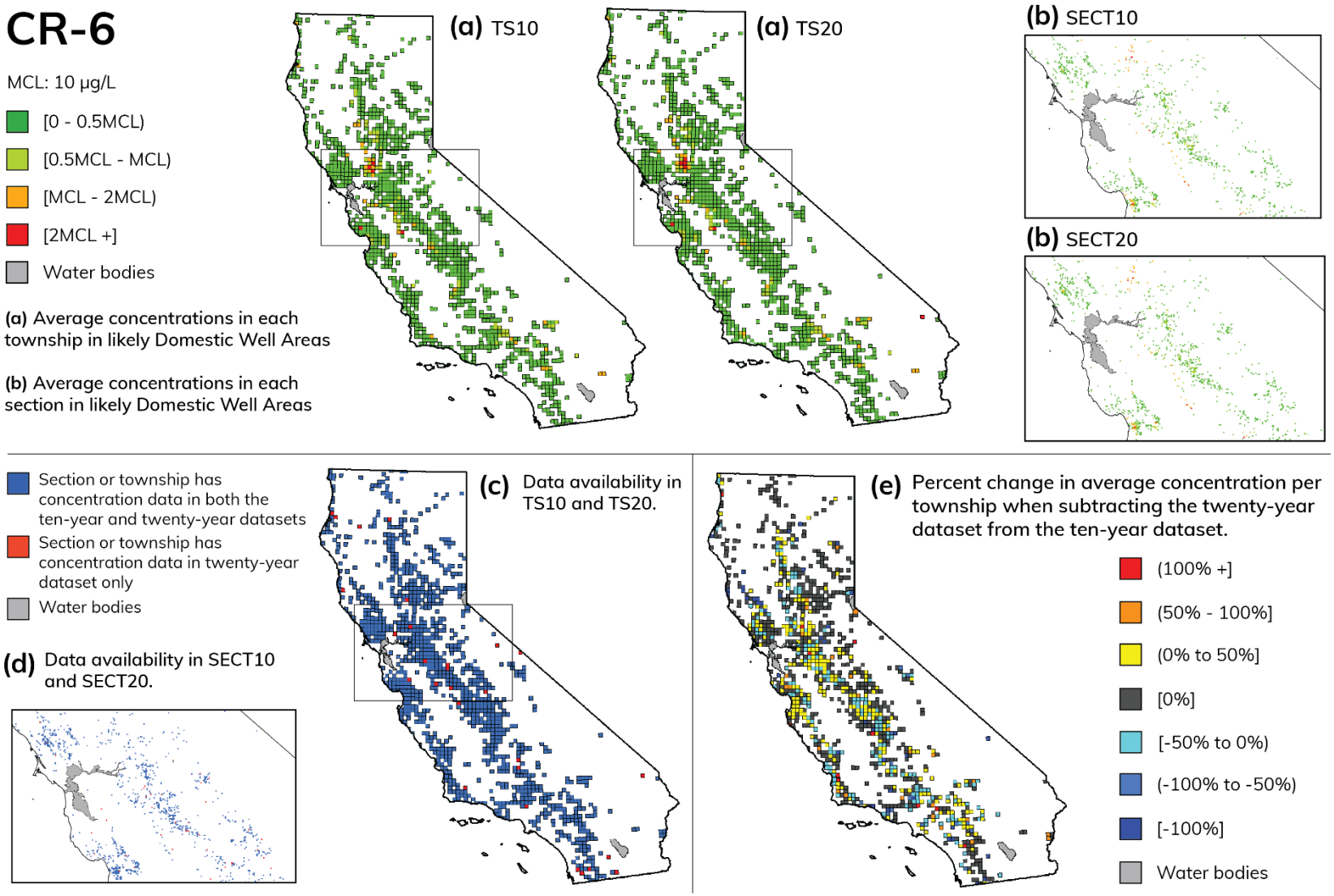

Figure 4: Maps displaying the results of the CR-6 models. 3a: Average concentration in each township containing a likely DWA. 3b: Average concentration in each section containing a likely DWA. 3c: Depiction of townships that have concentration data in the TS10 and TS20 models. 3d: Depiction of sections that have concentration data in the SECT10 and SECT20 models. 3e: Percent change in township concentration when comparing TS20 to TS10 estimates. E.g. An increase of 20% (yellow) indicates that when removing 2000-2010 sampling data for the ten-years dataset, the average concentration increased by 20%.

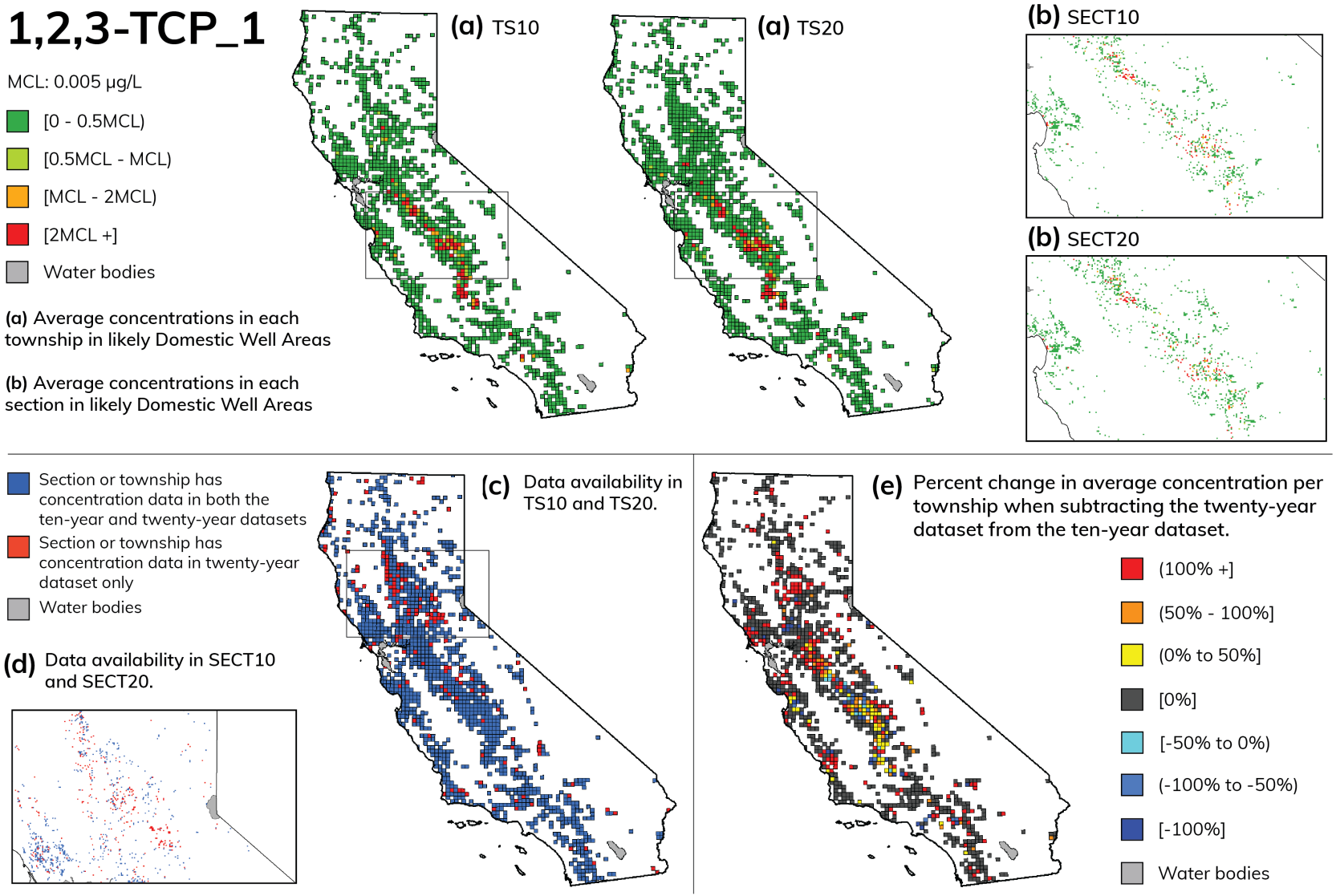

Figure 5: Maps displaying the results of the 1,2,3-TCP_1 models. 3a: Average concentration in each township containing a likely DWA. 3b: Average concentration in each section containing a likely DWA. 3c: Depiction of townships that have concentration data in the TS10 and TS20 models. 3d: Depiction of sections that have concentration data in the SECT10 and SECT20 models. 3e: Percent change in township concentration when comparing TS20 to TS10 estimates. E.g. An increase of 20% (yellow) indicates that when removing 2000-2010 sampling data for the ten-years dataset, the average concentration increased by 20%.

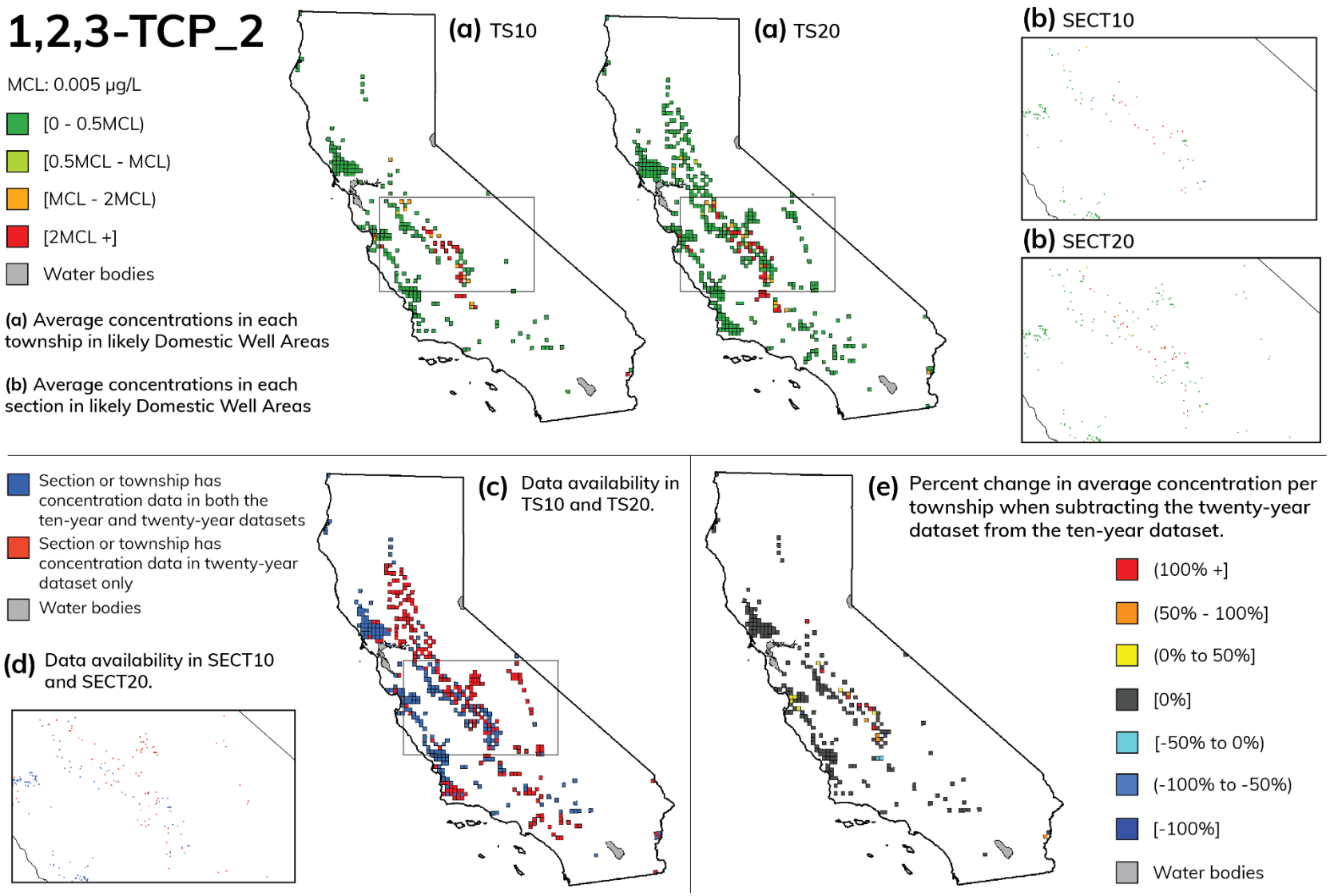

Figure 6: Maps displaying the results of the 1,2,3-TCP_2 models. 3a: Average concentration in each township containing a likely DWA. 3b: Average concentration in each section containing a likely DWA. 3c: Depiction of townships that have concentration data in the TS10 and TS20 models. 3d: Depiction of sections that have concentration data in the SECT10 and SECT20 models. 3e: Percent change in township concentration when comparing TS20 to TS10 estimates. E.g. An increase of 20% (yellow) indicates that when removing 2000-2010 sampling data for the ten-years dataset, the average concentration increased by 20%.

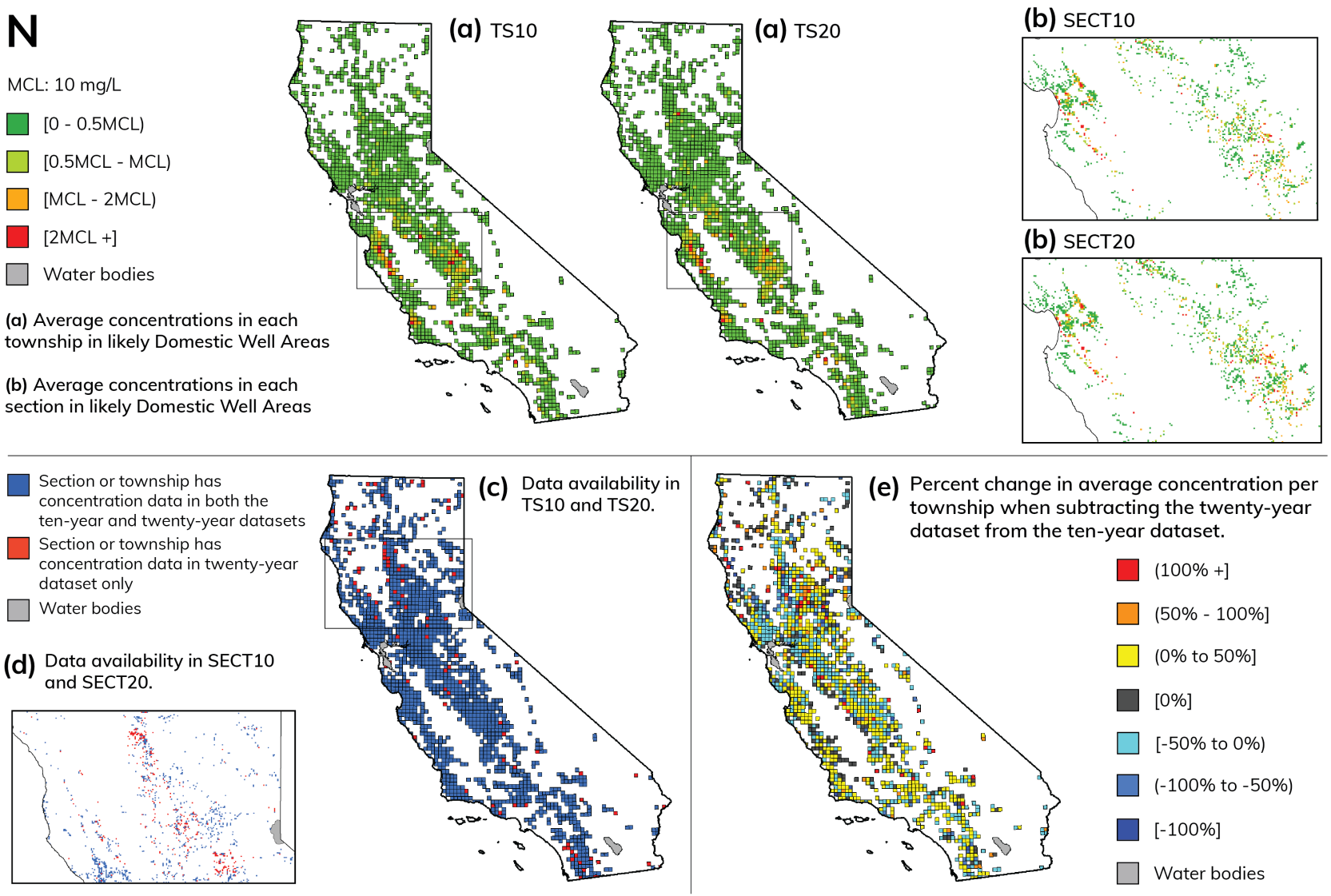

Figure 7: Maps displaying the results of the N models. 3a: Average concentration in each township containing a likely DWA. 3b: Average concentration in each section containing a likely DWA. 3c: Depiction of townships that have concentration data in the TS10 and TS20 models. 3d: Depiction of sections that have concentration data in the SECT10 and SECT20 models. 3e: Percent change in township concentration when comparing TS20 to TS10 estimates. E.g. An increase of 20% (yellow) indicates that when removing 2000-2010 sampling data for the ten-years dataset, the average concentration increased by 20%.

Temporal trends

As predicted, limiting the date range to a more recent time period lead to a decrease in the number of townships and sections with concentration data in likely DWAs for all contaminants (Table 2). For As, 1,2,3-TCP_1, and N, these decreases were most clearly observed in northern California (Figures 3c, 3d, 5c, 5d, 7c, 7d). For CR-6, the small percent decreases in number of sections (-5.11%) and townships (-2.54%) with concentration data were observed at the edges of the likely DWAs (Figures 4c, 4d). For 1,2,3-TCP_2, the large percent decreases in sections (-81.51%) and townships (-97.78%) with concentration data were observed throughout central California (Figures 6c, 6d). The percent increases in number of geographic units with concentration data was greater in sections than in townships for all contaminants.

The decreases in townships and sections with concentration data in likely DWAs did not lead to a decrease in the percentage of townships or sections with concentrations at or exceeding the MCL for most contaminants. For the two methods for 1,2,3-TCP at the section and township levels and for As at the section level, the percentage of townships or sections with concentrations at or exceeding the MCL increased (As: 1.41%; 1,2,3-TCP_1: 19.42%, 10.22%; 1,2,3,-TCP_2: 16.31%, 31.19%). In fact, 1,2,3-TCP had changes in percentage of townships or sections with concentrations at or exceeding the MCL that were one order of magnitude greater than that of the other contaminants. This indicates that for 1,2,3-TCP, using older data (2000-2010) lead to sizable decreases in the percentage of townships or sections with available concentration data that were out-of-compliance.

However, these decreases did not correspond to consistent effects in the estimates of at-risk people. For 1,2,3-TCP_1, the percentage decreased in the section comparison (-18.63%) and increased in the township comparison (13.67%). As, 1,2,3-TCP_2, abd N all demonstrated decreases in at-risk population estimates, meaning that excluding older data lead to fewer estimated at-risk people. The section comparison, with the exception of N, resulted in a larger magnitude of percent change in estimated at-risk population than in the township comparison for these contaminants.

Table 2: Percent change in the descriptive statistics for the temporal model comparison. These statistics describe the change in model outcomes from using a larger dataset including older data (twenty-year dataset, 2000-2019) to using a smaller dataset filtered to more recent data (ten-year dataset, 2010-2019). For example, the -27.55% change in number of sections with concentration data for the As section comparison should be interpreted as follows: The number of sections containing contaminant data for arsenic is 27.55% less when using a dataset of more recent water quality measurements (SECT10) instead of using a dataset including older water quality measurements (SECT20).

Contaminant (MCL)	Model comparison	%Δ in number of townships or sections with concentration data	%Δ in percentage of townships or sections with concentrations ≥ MCL	%Δ in population in areas with concentrations ≥ MCL	%Δ in percentage of total area with concentrations ≥ MCL
As (10 µg/L)	SECT20 : SECT10	-27.55%	1.41%	-28.20%	0.96%
	TS20 : TS10	-9.78%	-0.18%	-12.11%	-1.22%
CR-6 (10 µg/L)	SECT20 : SECT10	-5.11%	0%	1.31%	0.24%
	TS20 : TS10	-2.54%	-2.68%	0.93%	-2.52%
1,2,3-TCP_1 (0.005 µg/L)	SECT20 : SECT10	-28.17%	19.42%	-18.63%	19.50%
	TS20 : TS10	-12.91%	10.22%	13.67%	9.91%
1,2,3-TCP_2 (0.005 µg/L)	SECT20 : SECT10	-81.51%	16.31%	-106.63%	16.34%
	TS20 : TS10	-97.78%	31.19%	-5.06%	31.16%
N (10 mg/L)	SECT20 : SECT10	-17.73%	-0.48%	-3.02%	-0.17%
	TS20 : TS10	-6.21%	-3.59%	-20.66%	-5.23%

Geographic trends

For all contaminants, aggregating over townships resulted in larger at-risk populations than aggregating over sections with the same temporal dataset (Table 2). Although total area with concentrations equal to or exceeding the MCL increased comparing the section models to the township models for all contaminants, the change in percentage of total area with concentrations equal to or exceeding the MCL did not always increase. Percent changes decreased for CR-6, 1,2,3-TCP_1, and N, but increased for As and 1,2,3-TCP_2. For As, at-risk populations were mainly located in the Central Valley (Figures 3a, 3b). The at-risk populations for CR-6 were concentrated in Solano and Yolo counties (Figures 4a, 4b). Both methods for 1,2,3-TCP modeling suggested that the at-risk populations were located primarily in the San Joaquin Valley (Figures 5a, 5b, 6a, 6b). For N, the at-risk populations were located over a wider area in the Central Valley and the Central Coast (Figures 7a, 7b).

Table 3: Percent change in the descriptive statistics for the geographic model comparison. These statistics describe the change in model outcomes from aggregating over a smaller geographic unit (sections) to aggregating over a larger geographic unit (townships). For example, the 108.53% change in population in areas with concentrations at or exceeding the MCL for the As section comparison should be interpreted as follows: The number of estimated at-risk people is 108.53% greater when aggregating water well quality measurements for arsenic over larger areas (townships) instead of aggregating over smaller areas (sections).

Contaminant (MCL)	Model comparison	%Δ in population in areas with concentrations ≥ MCL	%Δ in percentage of total area with concentrations ≥ MCL
As (10 µg/L)	SECT10 : TS10	108.53%	12.37%
	SECT20 : TS20	82.35%	14.83%
CR-6 (10 µg/L)	SECT10 : TS10	49.25%	-16.80%
	SECT20 : TS20	49.83%	-14.49%
1,2,3-TCP_1 (0.005 µg/L)	SECT10 : TS10	457.37%	-13.28%
	SECT20 : TS20	305.63%	-2.96%
1,2,3-TCP_2 (0.005 µg/L)	SECT10 : TS10	1763.89%	41.54%
	SECT20 : TS20	3565.78%	16.47%
N (10 mg/L)	SECT10 : TS10	22.92%	-53.66%
	SECT20 : TS20	43.97%	-51.32%

DISCUSSION

This study aimed at evaluating whether different choices in water quality modeling affected estimates of at-risk areas and populations. I varied two modeling parameters: time and geography. I compared models using a dataset limited to ten years of water sampling data (2010-2019) to models using a dataset including twenty years of data (2000-2019). I also compared models aggregating water concentration data over sections to models aggregating over townships. The goal of this research was to evaluate the impacts of common water quality modeling choices made by various research groups, including Sacramento State's Office of Water Programs ("California Groundwater Risk Index (GRID)" 2018), Cal EPA's Office of Environmental Health Hazard Assessment ("CalEnviroScreen 3.0" 2016), UC Berkeley's Water Equity Science Shop (WESS, "The Drinking Water Tool, 2019" 2019), and the California State Water Resources Control Board (Methodology to Estimate Groundwater Quality Accessed by Domestic Wells in California 2019).

Implications

The results of this study demonstrate that water quality modeling choices regarding space and time have large impacts on estimates of at-risk areas and populations. As predicted, a larger data range resulted in improved geographic coverage of the state for all contaminants, but not without impacting the final model estimates. For the majority of the contaminants, increasing the date range lead to increases in the number of estimated at-risk people. However, changing the spatial resolution of the geographic unit over which concentration data was aggregated impacted estimates of at-risk populations to a greater extent. The magnitude of percent changes in population estimates exceeded 100% once in the time-variable analysis, while five of the geography-variable comparisons exceeded 100% change in magnitude. This finding suggests that although aggregating data over larger geographic units results in a more continuous model with values assigned over larger areas, at-risk population estimates will also inevitably increase.

Overall, the percent increases in number of geographic units with concentration data was greater in sections than in townships for all contaminants. This suggests that water wells gained when using the twenty-year datasets were more likely to be located in townships already containing wells from the ten-year datasets than in sections already containing wells from the ten-year datasets. The larger geographic footprint of townships likely lowers the probability of a water well gained in a twenty-year dataset intersecting a township without any water wells in the ten-year dataset.

Different spatial patterns of MCL violations arose for the four contaminants, which is in part due to where sources of contamination are located and where consistent water sampling is performed. The N models benefited from additional water sampling data provided by specialized programs like the Irrigated Lands Project ("Irrigated Lands Regulatory Program" 2019). Other contaminants had far less data. Coordinated efforts at the state level to bolster databases for contaminants like CR-6 and 1,2,3-TCP would improve the accuracy of future modeling efforts. In this study, I assumed that using data from sources other than the domestic wells dataset from GeoTracker GAMA, like pre-treatment Public Water System sampling data, still represented the quality of water accessed by domestic water wells. Other groups, like the GAMA group, decided to employ additional data cleaning methods on these datasets such as depth filtration (*Methodology to Estimate Groundwater Quality Accessed by Domestic Wells in California* 2019). These

additional data filtering methods further reduce the amount of data available to input into a model. In the absence of more frequent and extensive water quality monitoring from domestic water wells, it is impossible to say if one model is more "right" and estimates water quality more accurately than other models. Considering this, future modeling efforts should include sensitivity analyses that acknowledge and evaluate the impacts that different key choices have on model outcomes.

Data availability and 1,2,3-TCP

The case of 1,2,3-TCP illustrates how technological limitations in water contamination detection can impact future quality estimates and policy decisions. In the approach taken with the 1,2,3-TCP_1 models, any water quality measurements in both the ten-year and twenty-year datasets for which the reporting limit was missing were included in the final datasets. By contrast, the water quality measurements for which the reporting limit was missing were removed from the final datasets for the 1,2,3-TCP_2 models. The GAMA group recommended the second approach due to the unverifiable nature of the samples recorded without reporting limits (*Methodology to Estimate Groundwater Quality Accessed by Domestic Wells in California* 2019). This choice resulted in a drastic reduction in geographic coverage at both the section and township levels. For example, the difference between the number of sections with concentration data in 1,2,3-TCP_1 SECT20 model and the 1,2,3-TCP_2 SECT20 model was 2,430 sections (Table 1).

Notably, the estimates of at-risk populations varied on the magnitude of tens of thousands of people between these two approaches. At its greatest estimation difference, 87,032 people were estimated to be at-risk in the 1,2,3-TCP_1 TS10 model that were not predicted to be at-risk in the 1,2,3-TCP_2 TS10 model (Table 1). The results of the 1,2,3-TCP_1 and 1,2,3-TCP_2 models further support the assertion that modeling choices, including data cleaning choices, greatly impact model outcomes. The case of 1,2,3-TCP also highlights the need for more technologically reliable and geographically consistent water quality testing. Encouragingly, new methods such as in situ chemical reduction and in situ bioremediation appear promising for 1,2,3-TCP treatment, treatment methods are only as good as the methods used to identify areas of contamination (Merrill et al. 2019).

Future directions

Like with most research, this study dealt with an imperfect dataset curated throughout time by different organizations and people. The problematic nature of the 1,2,3-TCP data calls for instrumentation upgrades for future sampling efforts and for better training on how to determine and record reporting limits. Without reporting limits, it is difficult to determine if sampling data is in fact accurate and if it should truly be recorded as a value above or below the MCL. It is important to note also that the MCL is a convenient threshold that can be used to develop public health goals and track progress. MCLs will continue to change as public health research, detection ability, and cost of treatment change. It would be remiss to say that a community served by drinking water that is 0.0001% below the MCL is "safe" from any health effects. It is important to remind ourselves what modeling really does for us: Modeling helps us take a complicated system and simplify it so as to better understand it. The models produced in this and previous studies are not a literal depiction of water quality conditions in every community throughout California. Instead, models allow us to identify problematic areas to focus on. It is important, especially when modeling advises funding allocations for improvement projects, to still evaluate proposals on a community-by-community basis. That way, researchers best serve the people who have been institutionally refused clean, safe, and affordable drinking water and aid in achieving the Human Right to Water for everyone in California.

ACKNOWLEDGEMENTS

I would like to thank the ESPM 175 faculty – Patina Mendez, Samuel Evans, Leslie McGinnis, and Jessica Heiges – and fellow students for their support and perseverance through this challenging academic year. I would also like to thank my mentors, Clare Pace and Rachel Morello-Frosch, for their unwavering guidance and patience. Finally, I would like to thank my parents, Natalie and Lewis Krumm, for their interest in my research, their encouragement, and their love.

REFERENCES

2017 Annual Compliance Report. 2018. California Water Boards.

Balazs, C., R. Morello-Frosch, A. Hubbard, and I. Ray. 2011. Social Disparities in Nitrate-Contaminated Drinking Water in California's San Joaquin Valley. Environmental Health Perspectives 119:1272–1278.

Balazs, C., R. Morello-Frosch, A. Hubbard, and I. Ray. 2012. Environmental justice implications of arsenic contamination in California's San Joaquin Valley: a cross-sectional, cluster-design examining exposure and compliance in community drinking water systems. Environmental Health: 84.

BLM National Public Land Survey System. (n.d.). Cadastral/BLM_Natl_PLSS_CadNSDI (MapServer.) https://gis.blm.gov/arcgis/rest/services/Cadastral/BLM_Natl_PLSS_CadNSDI/MapServer.

Cadastral/BLM_Natl_PLSS_CadNSDI (MapServer). (n.d.). https://gis.blm.gov/arcgis/rest/services/Cadastral/BLM_Natl_PLSS_CadNSDI/MapServer.

CalEnviroScreen 3.0. 2016, December 29. https://oehha.ca.gov/calenviroscreen/report/calenviroscreen-30.

California Groundwater Risk Index (GRID). 2018. https://www.owp.csus.edu/grid/.

Carpenter, A., and M. Wagner. 2019. Environmental justice in the oil refinery industry: A panel analysis across United States counties. Ecological Economics 159:101–109.

Chemicals and Contaminants in Drinking Water | California State Water Resources Control Board. (n.d.). https://www.waterboards.ca.gov/drinking_water/certlic/drinkingwater/Chemicalcontaminants.html.

Cotton, M. 2018. Environmental Justice as Scalar Parity: Lessons From Nuclear Waste Management. Social Justice Research 31:238–259.

ESRI 2020. ArcGIS Desktop: Release 10.8. Redlands, CA: Environmental Systems Research Institute.

Faust, J., L. August, K. Bangia, V. Galaviz, J. Leichty, S. Prasad, R. Schmitz, A. Slocombe, R. Welling, W. Wieland, and L. Zeise. 2017. CalEnviroScreen 3.0. Office of Environmental Health Hazard Assessment.

GAMA Groundwater. 2019. https://gamagroundwater.waterboards.ca.gov/gama/gamamap/public/Default.asp.

Human Right to Water. 2019. https://www.waterboards.ca.gov/water_issues/programs/hr2w/.

Infrastructure. 2019. http://water.ca.gov/What-We-Do/Infrastructure.

Irrigated Lands Regulatory Program. 2019, October 21. https://www.waterboards.ca.gov/centralvalley/water_issues/irrigated_lands/.

Lee, C. 1987. Toxic wastes and race in the United States: a national report on the racial and socio-economic characteristics of communities with hazardous waste sites. The Commission.

McDonald, Y. J., and N. E. Jones. 2018. Drinking Water Violations and Environmental Justice in the United States, 2011–2015. American Journal of Public Health 108:1401–1407.

Meenar, M., J. P. Howell, and J. Hachadorian. 2019. Economic, ecological, and equity dimensions of brownfield redevelopment plans for environmental justice communities in the USA. Local Environment 24:901–915.

Methodology to Estimate Groundwater Quality Accessed by Domestic Wells in California. 2019. Whitepaper, State Water Resources Control Board.

Merrill, J.P., Suchomel, E.J., Varadhan, S. et al. Development and Validation of Technologies for Remediation of 1,2,3-Trichloropropane in Groundwater. Curr Pollution Rep 5, 228–237 (2019). https://doi.org/10.1007/s40726-019-00122-7

Monning, Bloom, and Garcia. 2019. Drinking Water.

Nitrates and Nitrites in Drinking Water | California State Water Quality Control Board. 2019. https://www.waterboards.ca.gov/drinking_water/certlic/drinkingwater/Nitrate.html.

Pace, C., C. Balazs, L. Cushing, and R. Morello-Frosch. 2019. UC Berkeley Water Equity Science Shop Domestic Well Communities Version 1.0, 2019.

Patrick, M. J., G. J. Syme, and P. Horwitz. 2014. How reframing a water management issue across scales and levels impacts on perceptions of justice and injustice. Journal of Hydrology 519:2475–2482.

R Core Team (2019). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

Real Estate Analysis & Mapping Application | LandVision™. (n.d.). https://www.digmap.com/platform/landvision/.

Schaider, L. A., L. Swetschinski, C. Campbell, and R. A. Rudel. 2019. Environmental justice and drinking water quality: are there socioeconomic disparities in nitrate levels in U.S. drinking water? Environmental Health: A Global Access Science Source 18:1–15.

State Water Resources Control Board. 2015, October 12. https://www.waterboards.ca.gov/water_issues/programs/ust/electronic_submittal/about.shtml.

TIGER/Line Shapefiles. (n.d.). https://www.census.gov/geographies/mapping-files/time-series/geo/tiger-line-file.html.

Tracking California, Public Health Institute. Water Boundary Tool. Accessed 4/7/2020 from www.trackingcalifornia.org/water

Water Systems --- Tracking California. (n.d.). https://trackingcalifornia.org/water-systems/water-systems-landing.

Well Completion Reports. (n.d.). http://water.ca.gov/Programs/Groundwater-Management/Wells/Well-Completion-Reports.

Wikstrom, K., T. Miller, H. E. Campbell, and M. Tschudi. 2019. Environmental Inequities and Water Policy During a Drought: Burdened Communities, Minority Residents, and Cutback Assignments. Review of Policy Research 36:4–27.

APPENDIX A: Data Source Selection

GeoTracker GAMA contains the following datasets: Department of Pesticide Regulation; Department of Water Resources; GAMA – Domestic Wells; GAMA – Special Studies; GAMA – Priority Basin Project; Local Groundwater Projects; Monitoring Wells (Water Board Regulated Sites); Public Water System Wells; and National Water Information System ("GAMA Groundwater" 2019).

Data availability from different sources

The availability of data is inconsistent between datasets depending on the contaminant. For example, the Department of Pesticide Regulation does not have data for As, 1,2,3-TCP, CR-6 or N, and therefore it is not a data source used in this analysis. Out of the four contaminants included in this study, N has the most observations in terms of both the number of datasets and the number of unique water wells with available data. In contrast, CR-6 has the least representation of the four contaminants.

GeoTracker GAMA water well types

Although the data in GeoTracker GAMA are measurements of groundwater quality, the measurements are taken from different types of wells. The wells measured in the GAMA – Domestic Wells dataset are strictly domestic water wells. The Public Water System (PWS) Wells dataset is the largest dataset in terms of number of unique wells measured and geographic coverage. The measurements in the PWS dataset are raw groundwater quality measurements, suggesting that they are representative of the aquifer quality from which nearby domestic wells source their water. I included all water wells regardless of depth even though PWS wells tend to be deeper than domestic wells (Faust et al. 2017). The Department of Water Resources, GAMA – Special Studies, GAMA – Priority Basin Project, and Local Groundwater Projects datasets represent measurements from a mixture of domestic, public, irrigation, and monitoring wells.

Monitoring wells: Data exclusion rationale

The wells measured in the Monitoring Wells (Water Board Regulated Sites) dataset typically represent water wells that are installed on industrial land-use sites ("State Water Resources Control Board" 2015). For instance, there are sub-categories in the dataset that distinguish water wells installed on Leaking Underground Storage Tanks cleanup sites, land disposal sites, military sites, and oil and gas development sites ("State Water Resources Control Board" 2015). The subset of the Monitoring Wells dataset that includes wells located in residential areas is the Irrigated Lands Program ("Irrigated Lands Regulatory Program" 2019).

Selected data sources

The goal of this research is to elucidate the implications of model choices when using environmental data in a public health context. I selected the following eight GeoTracker GAMA datasets because they are representative of groundwater that could potentially supply domestic wells: Department of Water Resources; GAMA – Domestic Wells; GAMA – Special Studies; GAMA – Priority Basin Project; Local Groundwater Projects; Irrigated Lands Projects (a subset of the Monitoring Wells dataset); Public Water System Wells; and National Water Information System (NWIS).

APPENDIX B: Data cleaning

The goal of this data cleaning was to standardize the data within and between datasets. I cleaned the data using three columns in GeoTracker GAMA datasets: the contaminant concentration (RESULTS), qualifier (QUALIFER), and reporting limit (RL). In the first stage, I cleaned the QUALIFER column (Table B1). If there were any rows for which the RESULTS and RL columns were missing ('NA' or 'UNK') and the QUALIFER was an equal sign ('='), the row was removed because it lacked any useful data on the contaminant concentration at the time of sampling. If a number was provided for the RESULTS and RL columns, then the QUALIFER column was cleaned so that only two QUALIFER values remained ('<' and '='). This methodology is in line with that of the State Water Resources Control Board's (*Methodology to Estimate Groundwater Quality Accessed by Domestic Wells in California* 2019 p. A-1)

In the second stage, I cleaned the RESULTS column. If for any rows the result was either missing a value or zero ('NA' or '0'), the qualifier was a less than sign ('<'), and the reporting limit was missing a value ('NA'), the result was assigned a zero. If for any rows the result was a number but the qualifier was a less than sign ('<') and the reporting limit was missing or a number ('NA' or Number), the assumption was that the result was below the reporting limit and was therefore assigned a zero. If the qualifier was an equals sign ('='), the result was kept as reported.

Table B1: Data cleaning decisions according to the results (RESULTS), qualifier (QUALIFER), and reporting limit (RL) columns in the GeoTracker GAMA datasets. The bolded entries indicate the exact value in the dataset. The non-bolded entry (Number) represents any positive decimal.

Stage	Results (RESULTS)	Qualifier (QUALIFER)	Reporting Limit (RL)	Action
1	**NA**	=	**NA**	Remove the row.
	Number	**ND**, <	Number	Assign < to QUALIFER.
	Number	**NA**, **I**, **Q**, **S**, **-**	Number	Assign = to QUALIFER.
2	**NA**	<	**NA**	Assign 0 to RESULTS.
	0	<	**NA**	Numerical entry (0) is taken as RESULTS.
	Number	<	Number	Assign 0 to RESULTS.
	Number	=	Number	Numerical entry taken as RESULTS.

Reporting limit less than the MCL

In some cases, the reporting limit (RL) exceeded the MCL but the recorded contaminant concentration (RESULTS) was less than the reporting limit. For example, an As measurement could have a recorded concentration of 5 µg/L and a reporting limit of 15 µg/L. The MCL for As is 10 µg/L, meaning that the reporting limit for this measurement exceeded the MCL. Because the concentration is less than the reporting limit, it is unclear whether 5 µg/L is an accurate result. It is possible that the actual result was above the MCL of 10 µg/L (between 10 µg/L and 15 µg/L). Therefore, I dropped all rows for which there was a numeric reporting limit greater than the contaminant's MCL and a concentration less than the reporting limit. If the reporting limit was greater than the contaminant's MCL but the recorded concentration was greater than the reporting limit, the value was kept as the result.

Missing reporting limits and 1,2,3-TCP

All of the contaminants' ten-year and twenty-year datasets contained rows with missing reporting limits. For these rows, the same cleaning methods detailed above could not be applied. For As, CR-6, and N, the year-average reporting limits were above the MCLs I used in this study (10 µg/L, 10 µg/L, and 10 mg/L, respectively). Because the average reporting limits were above these contaminants' MCLs, I kept all of the rows for which the reporting limit was missing.

For 1,2,3-TCP, however, the average reporting limit was above the MCL (0.005 µg/L) each year from 2000 to 2019. This indicates that there were serious limitations in monitoring technology for 1,2,3-TCP, and that many of recorded concentrations could be inaccurate. In order to understand the effects of monitoring limitations on modeling outcomes, I produced two sets of the 1,2,3-TCP datasets (1,2,3-TCP_1 and 1,2,3-TCP_2). For the first approach (1,2,3-TCP_1), I included the recorded concentrations with missing reporting limits ('UNK'). For the second approach (1,2,3-TCP_2), I removed all rows for which the reporting limit was missing. Therefore, there are eight total models for 1,2,3-TCP representing two different data cleaning approaches in this study.