International Journal of Geo-Information

MDPI

Article

Analysis of Groundwater Nitrate Contamination in the Central Valley: Comparison of the Geodetector Method, Principal Component Analysis and Geographically Weighted Regression

Anil Shrestha * and Wei Luo

Geographic and Atmospheric Sciences, Northern Illinois University, DeKalb, IL 60115, USA; wluo@niu.edu
* Correspondence: ashrestha1@niu.edu; Tel.: +1-815-517-6485

Received: 28 July 2017; Accepted: 18 September 2017; Published: 26 September 2017

Abstract: Groundwater nitrate contamination in the Central Valley (CV) aquifer of California is a ubiquitous groundwater problem found in various parts of the valley. Heavy irrigation and application of fertilizer over the last several decades have caused groundwater nitrate contamination in several domestic, public and monitoring wells in the CV above EPA's Maximum Contamination level of 10 mg/L. Source variables, aquifer susceptibility and geochemical variables could affect the contamination rate and groundwater quality in the aquifer. A comparative study was conducted using Geodetector (GED), Principal Component Analysis (PCA) and Geographically Weighted Regression (GWR) to observe which method is most effective at revealing environmental variables that control groundwater nitrate concentration. The GED method detected precipitation, fertilizer, elevation, manure and clay as statistically significant variables. Watersheds with percent of wells above 5 mg/L of nitrate were higher in San Joaquin and Tulare Basin compared to Sacramento Valley. PCA grouped cropland, fertilizer, manure and precipitation as a first principal component, suggesting similar construct of these variables and existence of data redundancy. The GWR model performed better than the OLS model, with lower corrected Akaike Information Criterion (AIC) values, and captured the spatial heterogeneity of fertilizer, precipitation and elevation for the percent of wells above 5 mg/L in the CV. Overall, the GED method was more effective than the PCA and GWR methods in determining the influence of explanatory variables on groundwater nitrate contamination.

Keywords: nitrate; groundwater; geodetector; Central Valley; statistics

1. Introduction

The Central Valley (CV) has been one of the most productive agricultural regions in the United States for the last 50 years, producing huge quantities of agricultural products. CV, with a total area of 47,000 km^2, occupies less than 1% of the total farmland in the United States and still produces 8% of the country's agricultural output in cash value (Figure 1). The valley accounts for about one-sixth of the irrigated land and one-eighth of the groundwater pumpage in the United States [1]. Approximately 75% of the irrigated land in California and 17% of the nation's irrigated land is in the CV [2]. However, the large percent of agricultural land has come at a cost of heavy pumping of groundwater from wells for irrigation [3] and groundwater nitrate contamination from fertilizer application [4]. Groundwater withdrawals from the CV principal aquifer is 13% of the country's total withdrawal, making it the second largest in the United States [5]. Heavy withdrawal of groundwater and surface water diversion for irrigation have changed the groundwater flow pattern of the CV, which when compounded with the heavy nitrogen fertilizer application rate, has complicated the groundwater contamination by nitrate in the valley [6].

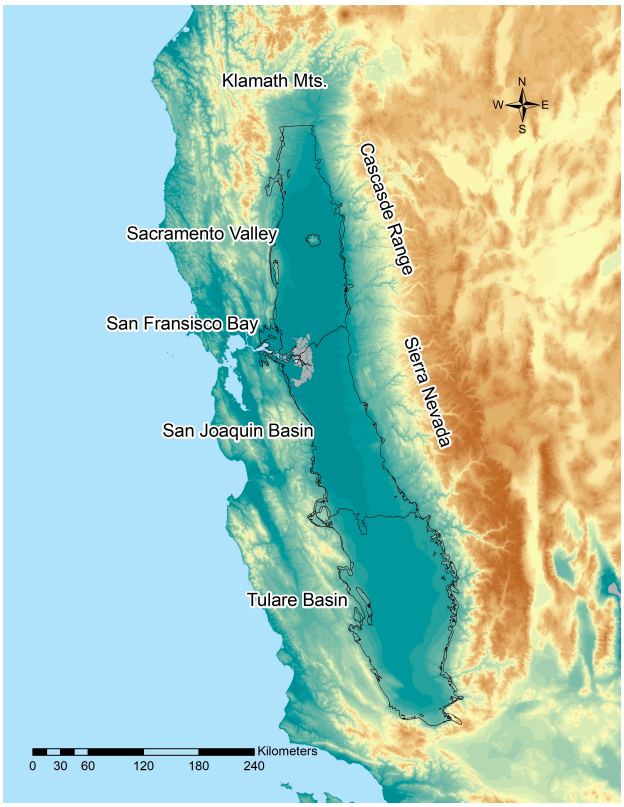

Figure 1. Central Valley Aquifer [7].

The USGS National Water-Quality Assessment (NAWQA) program estimated a nitrate contamination greater than the national median in San Joaquin and Tulare basin of the CV [8]. Counties in San Joaquin valley were detected above the Environmental Protection Agency(EPA)-maximum contaminant level (MCL) of nitrate in over 40% of total 200 domestic well samples [9]. Drinking water contaminated with nitrate can cause several health problems; some, like methemoglobinemia, can even cause death in children [10].

The application of fertilizer, manure, septic tanks and domestic animals are some of the major sources of nitrate contamination in the groundwater of the CV [11]. However, the concentration of nitrate in the aquifers can be altered by several other environmental factors like rainfall, soil conditions, permeability, and the geochemical condition of the valley. For example, groundwater in the central part of the CV has a more reducing condition that converts nitrate to nitrogen gas (denitrification), and therefore, nitrate in groundwater is lower compared to the eastern part of the valley. The alluvial deposit in the eastern part is relatively coarse grained and receives more rainfall, creating an oxic condition, whereas small grained sediments and the long residence time of groundwater creates a reducing condition as it flows towards the axis of the valley [12,13]. Rainfall is comparatively higher in northern Sacramento Valley than the southern Tulare Basin and San Joaquin Basin in the CV, which can dilute the concentration of nitrate or percolate dissolved nitrate into deeper parts of the aquifer. In the CV, the hydrogeological characteristics of San Joaquin Valley have already been impacted severely due to heavy pumping of groundwater, altering the groundwater flow pattern and causing most of the groundwater to flow towards the cone of depression [3]. This change in the groundwater flow pattern could alter the transport mechanism of the nitrate dissolved in the groundwater as well. Therefore, several of these environmental variables could make it difficult to predict the fate and transportation of this contaminant in the CV.

Several studies in the San Joaquin Basin, Sacramento Valley and other counties have already confirmed high levels of nitrate in the groundwater of the CV [14–18]. Statistical methods have been

used to study the relationship between groundwater nitrate contamination and predictor variables. For example, multivariate statistical methods like ordinary least square regression (OLS) have been applied in many studies to investigate the relationship between groundwater nitrate contamination and land use conditions [19–21]. However, OLS has a drawback of overfitting the data by regressing random error in the model rather than the relationship between the variables when there are numerous variables. The OLS makes several assumptions on the linearity, homoscedasticity, normality and multicollinearity of the data, which are often difficult to apply to nitrate data as they are often skewed. These assumptions lead to a selection of only a few variables based on the researcher's knowledge or use statistical methods that reduce the number of variables. Logistic regression is another common method applied to find the areas above a certain threshold of MCL or background concentration level of nitrate in groundwater [22–24]. However, logistic regression only calculates the probability of concentration above a certain threshold and does not calculate the direct concentration of contaminants.

Non-parametric tests like the Mann-Whitney test and Kruskal-Wallis test were applied to reveal differences in mean nitrate concentrations in well samples of different land-use types in the eastern San Joaquin Valley [25]. Another study conducted at the eastern San Joaquin Valley, using the Kruskal Wallis test and Wilcoxon rank-sum test, confirmed a higher nitrate concentration in shallow aquifers due to non-point sources [26]. The Regional Kendal test has also been applied to study the decadal trend of nitrate concentration in different physiographic sub-regions of the CV [13].

Recently, machine learning methods have been used to predict the groundwater nitrate contamination in the CV [27,28]. Nolan et al. applied a statistical learning framework to domestic wells data to optimize the over prediction of three machine learning methods: Artificial Neural Network (ANN), Boosted Regression Tree (BRT) and Bayesian Network (BN). Their results showed BRT had the highest predictive performance. These models were also compared to the random forest regression (RFR) model and multiple linear regression (MLR); the results showed that the BRT model was comparable with RFR, but MLR predicted poorly [28].

Principal Component Analysis (PCA) is commonly applied to study the hydrogeochemical parameters of groundwater, including nitrate. PCA reduces the dimensions of data by using a linear combination of variables to capture the maximum variance in the data [29–31]. The principal scores for each component can also be calculated, which are then used as the independent variables to build the predictive model [32–34] or compare it with other variables of interest [35]. The correlation matrix between the variables and the loadings of each variable in the principal components are used to investigate the contamination process in the study area. The grouping of redundant variables in the principal components removes the multicollinearity and reduces the number of independent variables into principal components that still explain most of the variance in the data [35]. Burrow et al. analyzed geochemical variables including nitrate and other physio-chemical factors using PCA in eastern San Joaquin Valley. The result showed that nitrate concentration were higher in coarse-grained sediments, soil with a low percent of clay and under oxidizing geochemical conditions [25].

Geographically Weighted Regression (GWR) has been commonly used in many studies to determine the effect of spatial heterogeneity on the explanatory variable. OLS assumes that relationships between explanatory variables and the outcome variable are homogeneous (or stationary) across the study area, but GWR assumes that the relationship varies across the study area (non-stationary). This method models the spatially varying relationships by generating individual regressions for each data point where nearby observations are given more weight. The method minimizes the spatial autocorrelation in the residuals and generates local coefficient maps to observe the spatial heterogeneity over the area [36]. The GWR method has shown that water quality indicators varied as the landuse changed in the watersheds [37–39]. Groundwater quantity was also found to change spatially depending on landuse type [40]. However, studies have also been performed showing multicollinearities between the coefficients obtained from the GWR model [41].

Geodetector (GED) is a relatively new statistical method that measures spatial stratified heterogeneity and tests its significance [42,43]. It can detect important contributing explanatory variables

as well as the risk areas of contamination. In addition, it also quantifies if a geographical stratum (associated with one suspected explanatory variable) is more significant than another geographical stratum (of another suspected variable). Also, the method analyses whether two determinants can interact to increase or decrease the chance of contamination in the aquifer [42,44]. The GED method has been applied to the analysis of groundwater nitrate contamination in the CV [45]. Here, we compare the results of the GED method and those of the PCA and GWR methods to better understand the contamination process and to illustrate the advantages and additional information that the GED method provides in understanding the groundwater contamination process in the CV. The specific objectives of this research are as follows: (a) To understand the controlling explanatory variables of groundwater nitrate contamination in the CV using three statistical methods: Geodetector (GED), Principal Component Analysis (PCA) and Geographically Weighted Regression (GWR); (b) To identify contributing geographical stratum or the risk areas of contamination in the CV; (c) To perform a comparative study of the GED method with PCA and GWR in understanding the contamination process.

2. Methods

The watershed was selected as the basic analysis unit to quantify the effects of explanatory variables on groundwater nitrate contamination as groundwater and surface water interact in hydrologic processes at a watershed scale. The 12-digit hydrologic unit watersheds of the CV (Figure 2) were selected to process the data. Watershed data were downloaded from USGS-National Standard for the Spatial Data Accuracy and Watershed Boundary Dataset. This is the finest resolution of hydrologic unit hierarchy. There are total of 656 watersheds in the CV.

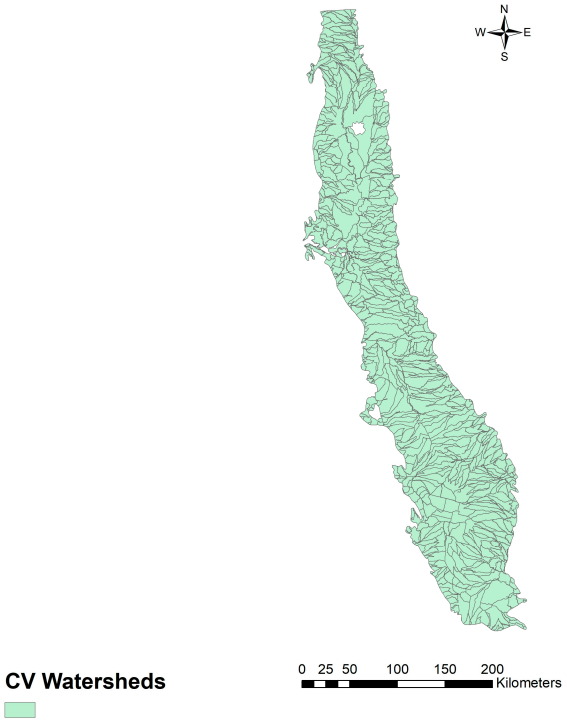

Figure 2. CV watersheds as the basic analysis unit.

Nitrate data were collected from the National-Water Quality Assessment (NAWQA) project; National Water-Quality Information System (NWIS); and the State Water Resource Control Board-Groundwater Ambient Monitoring and Assessment Program (GAMA). The data were downloaded from these websites for the period of 2002 to 2014. NAWQA and NWIS measured nitrate concentration as NO_3-N and GAMA as NO_3. To make the concentration data consistent, NO_3 was converted to NO_3-N. The dataset consists of a total of 2516 well samples spread throughout the Central Valley

(Figure 3). Multiple measurements of nitrate data for each well sample were averaged over the study period to obtain the temporal average for each well sample.

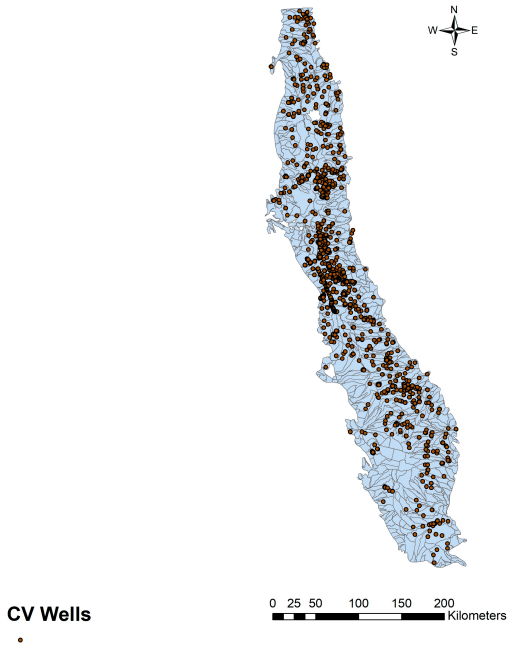

Figure 3. Well sampling locations for groundwater NO_3 concentration.

The explanatory variables and their sources are listed in Table 1. There are 12 different explanatory variables. Farm fertilizer, manure and cropland can be categorized as source variables as they directly contribute to the groundwater nitrate. Permeability, precipitation, slope, elevation and clay are the variables that determine the susceptibility of the aquifer to groundwater nitrate contamination. Dissolved oxygen, iron and manganese represent the geochemical condition of groundwater which determines the oxidation-reduction potential of groundwater that could affect the denitrification process of nitrate. Both nitrate data and all of the explanatory variables were processed at the watershed level in the CV. For each watershed, percent of wells above a mean nitrate concentration of 5 mg/L ($PW_{N>5}$) were calculated (Figure 4) in ArcGIS using the formula:

$$PW_{N>5} = \frac{\text{Number of wells exceeding temporal avg.} > 5 \text{ mg/L in a watershed}}{\text{Total number of wells in each watershed}} \quad (1)$$

Table 1. Explanatory variables and data sources.

Explanatory Variables	Data Sources
Farm Fertilizer (kg/ha)	United States Geological Survey (USGS) [46]
Manure (kg/ha)	United States Geological Survey (USGS) [47]
Cropland (%)	National Landcover Database (NLCD) [48]
Permeability(in/h)	STATSGO Soil Characteristics for the Conterminous United States [49]
Precipitation (mm)	PRISM Climate Data. [50]
Slope (%)	Elevation Derivatives for National Applications (EDNA) [51]
Elevation (m)	National Elevation Dataset (NED) [51]
Clay (%)	United States Geological Survey (USGS) [52]
Recharge Rate (mm/year)	United States Geological Survey (USGS) [53]
Dissolved Oxygen (mg/L)	NAWQA and NWIS [54,55]
Iron and Manganese (mg/L)	NAWQA and NWIS [54,55]

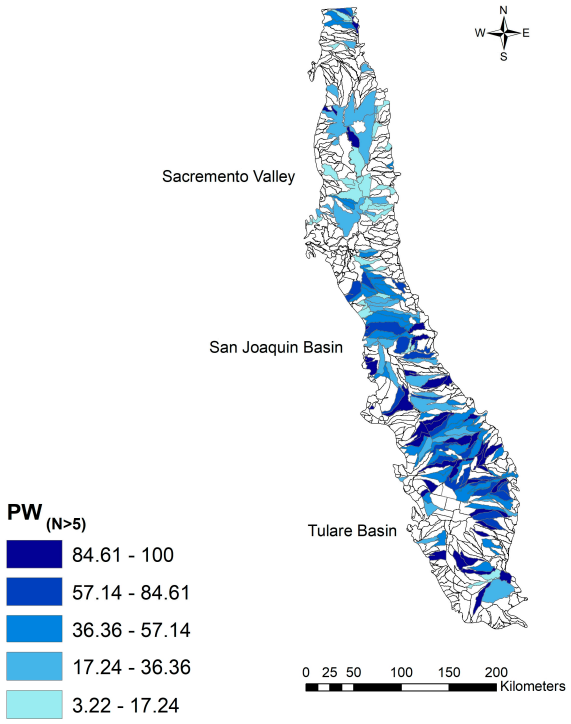

Figure 4. Watersheds with $PW_{N>5}$ mg/L.

A nitrate concentration of 5 mg/L was selected as the threshold level because previous studies have used concentrations above 5 mg/L as the background concentration level for nitrate [13]. Nitrate concentration in nature is usually less than 1 mg/L as Nitrate-N and therefore, concentrations above 5 mg/L can be attributed to other sources [56]. A total of 1014 well samples had a temporal average greater than 5 mg/L.

Explanatory variables were also processed at the watershed level to correspond with the $PW_{N>5}$ for each watershed. They were then reclassified into different zones to represent the spatially stratified heterogeneity of explanatory variables in the CV. Classification was performed using natural break classification as described in [57]. Details on the processing of explanatory variables can be found in [45]. The data set was then analyzed using three different methods: Geodetector (GED), Principal Component Analysis (PCA) and Geographically Weighted Regression (GWR).

2.1. Geodetector

The Geodetector (GED) method was applied using GeoDetector Software 2007 [42]. GED reveals the relationship between the spatial distribution of the nitrate contamination pattern and that of the potential environmental risk factors. The assumption is that if an environmental factor leads to nitrate contamination, the contamination would exhibit a spatial distribution similar to that of the environmental factor [42]. Instead of directly comparing the contamination rate with some measure of the environment factor (e.g., rainfall), it calculates the variance of the contamination rate within each zones of the environmental factor (e.g., each rainfall class) and compares the local variances of the contamination rate within each zone of the factor with the global variance of the contamination rate within the entire study area.

For example, in Figure 5a, we have watersheds in the CV used as the basic analysis unit for which we calculated the $PW_{N>5}$ for each watershed and used this as the response variable. Figure 5b shows the CV subdivided into different rainfall sub-regions (low to high) based on the amount of rainfall in the valley. If we overlay the watershed unit over the rainfall map (Figure 5c), we can calculate the mean and variance of $PW_{N>5}$ in each different rainfall sub-region. This variance of $PW_{N>5}$ in each

sub-region (local variance) is then compared with the variance of the entire CV (global variance) to calculate the Power of Determinant (PD) value. This can be represented by the formula:

$$\mathrm{PD} = 1 - \frac{A_{C1} \cdot Var_{C1} + A_{C2} \cdot Var_{C2} + \ldots + A_{C5} \cdot Var_{C5}}{A \cdot VarE} \quad (2)$$

where Var_{Ci} (i = 1, 2, . . . , 5) is the variance of $PW_{N>5}$ in each sub-region of rainfall class; $VarE$ is the variance of all the watersheds within the entire CV; A represents the area of the entire study area; A_{Ci} represents the area of each sub-region of rainfall class.

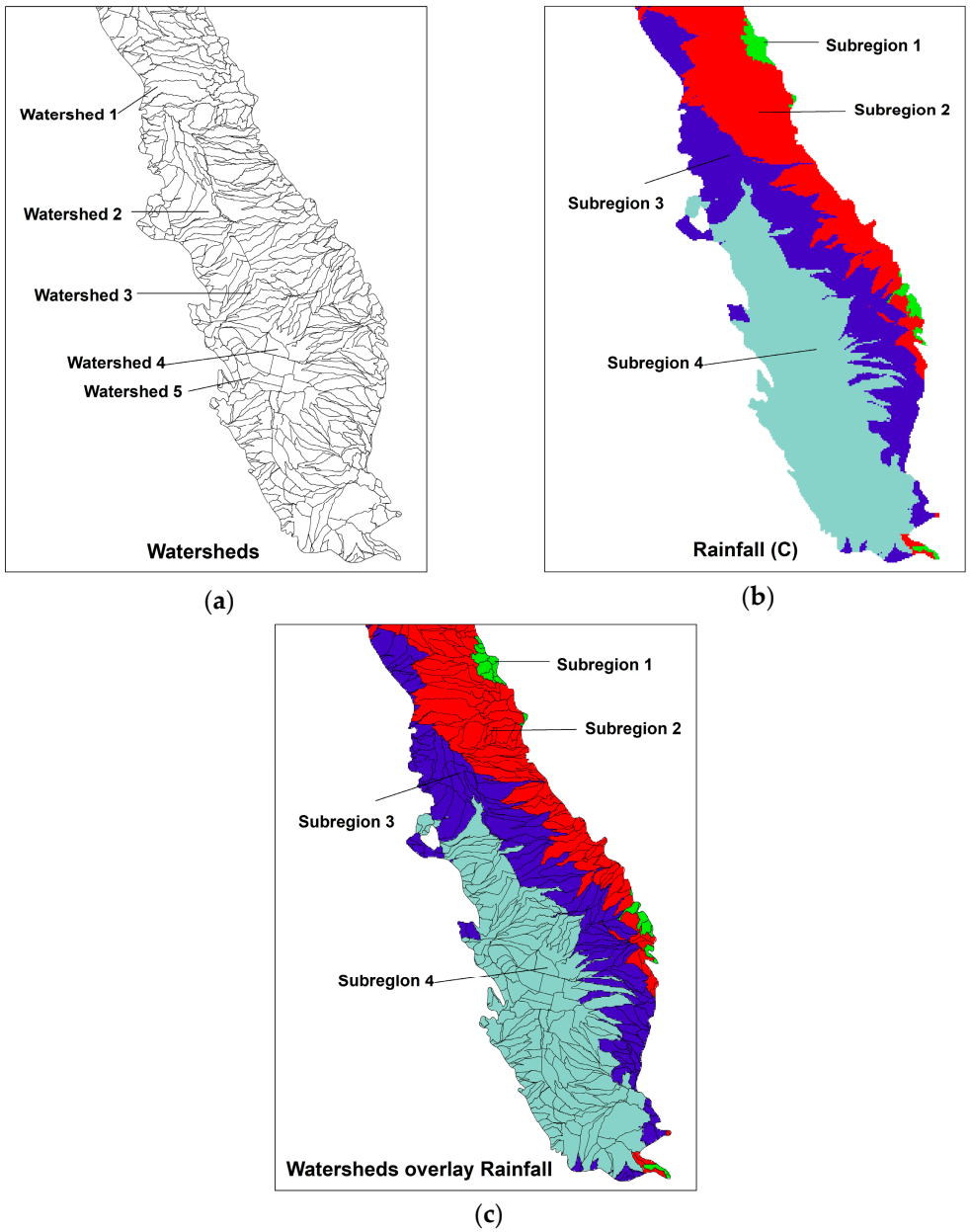

Figure 5. Illustration of the Geodetector method and basic analysis unit (watershed): (**a**) basic analysis unit watershed; (**b**) sub-regions of the study area based on one factor (rainfall); (**c**) overlay of the basic analysis unit and sub-regions for Geodetector analysis.

The value of PD ranges from 0 to 1. The PD value will be equal to 1 if the explanatory variable rainfall (C) completely controls the groundwater nitrate contaminations, and it will be zero if it is

completely independent of groundwater nitrate contaminations. Values closer to 1 suggest greater contribution to groundwater nitrate by rainfall as a determinant variable.

Risk detector calculates the mean value of $PW_{N>5}$ of each sub-region (M_{Ci}) where $i = 1, 2, \ldots, 5$ and tests if the mean value is statistically significantly different from the mean value of all the remaining sub-regions. The explanatory variable, rainfall (continuous variable) is discretized from Level 1 to Level 5 (low to high), and risk detector identifies which level has the highest mean nitrate concentration compared to others and if it is statistically significant.

Ecological Detector compares if the geographical stratum of C (e.g., rainfall) contributes to the groundwater nitrate contamination more significantly than another geographical stratum of D (e.g., Permeability). If C (e.g., rainfall) has more control over the contamination, it will have a lower dispersion variance (σ^2).

Interaction detector calculates the combined effect of two explanatory variables in the CV. Individual PD values of variables are then compared with the combined PD values of given variables to assess whether they strengthen, weaken or are independent of each other, e.g., nonlinear enhancement exists if the PD value of the new factor is greater than the PD value of each individual factor summed together [42]. More details of the method and statistical tests can be found in [42,43].

2.2. Principal Component Analysis

Principal Component Analysis (PCA) was used to analyze the data for all 12 explanatory variables using SPSS. PCA is a dimension reduction technique that reduced the 12 explanatory variables into a few number of principal components (PC) that explained most of the variance of the data. Eigenvalues based on the standardized correlation matrix of the data were utilized and rotated using varimax rotation to maximize the variation. This is achieved by rotating the original variables coordinate system into a new coordinate system (orthogonal to each other) towards the direction of maximum variance of the data. The first PC measured the maximum variance in the data set, the second PC measured less variance than the first PC and the third PC measured less than the second PC and so on. The contribution of each variable to the newly formed PC is called the loading. Higher loading of a variable in a PC means that variable explained most of the variance in that PC. These newly generated PCs are the uncorrelated variables derived from the correlated or redundant explanatory variables. All the assumptions of the PCA, i.e., that variables are continuous and linearly correlated, and adequate sampling size and outliers were checked before performing the test using SPSS [30,31].

2.3. Geographically Weighted Regression

Geographically Weighted Regression (GWR) was performed on the data using ArcGIS 10.4. A properly specified Ordinary Least Square (OLS) model was fit prior to the GWR analysis. All the assumptions of normality, multicollinearity and heteroskedasticity of the OLS model were met before performing GWR. The GWR model is a spatial regression model which estimates OLS-like regressions for each feature in the data set. The OLS model is a simple linear model where the relationship between dependent variable and explanatory variable is assumed constant over the study area and residuals are assumed to be independent. However, normal features that are nearby could be more related than those that are far away (Tobler's Law); hence, spatial autocorrelation exists if residuals are autocorrelated and the assumption of OLS is violated. GWR is an effective approach when there is spatial heterogeneity and the relationship between variables vary over the study area. The GWR model is therefore also known as a local model which considers the spatial non-stationarity in data by building a separate model for the explanatory variable and outcome variable. The key to analysis using the GWR model is defining a proper bandwidth or number of neighbors for each target feature. The shape and size of the bandwidth is dependent on user input for the kernel type, bandwidth method, distance, and number of features. Bandwidth or number of neighbors determines the degree of smoothing in the model. If the kernel type is fixed, the Gaussian kernel is used to solve each local regression analysis using a fixed distance. The adaptive kernel type adapts to the density of the

neighbors. Bandwidth method specifies the extent of the kernel. Akaike Information Criterion (AIC) or CV (Cross Validation) selects the optimal distance or number of neighbors [58].

3. Results

3.1. Geodetector Method

The results from the GED method are shown in Table 2. The PD values for source variables (fertilizer and manure) and aquifer susceptibility factors (precipitation, elevation, percent clay) were statistically significant variables with *p*-values less than 0.05. Geochemical variables (dissolved oxygen, iron and manganese) were not found to be statistically significant using this method. Risk Detector results (Table S1) showed geographical areas with low rainfall (San Joaquin and Tulare Basin) have higher $PW_{N>5}$ (Figure 6).

Table 2. PD values for explanatory variables.

Variables	PD	*p*-Value
Precipitation (PPT)	0.27	<0.01
Fertilizer (FERT)	0.21	<0.01
Elevation (ELVN)	0.18	<0.01
Manure (MANU)	0.16	0.01
Clay (CLAY)	0.10	0.03
Dissolved Oxygen (DO)	0.09	0.09
Permeability (PERM)	0.09	0.14
Iron (FE)	0.06	0.14
Slope (SLP)	0.03	0.21
Cropland (CROP)	0.07	0.22
Manganese (MN)	0.0	1.00
Recharge Rate (RECH)	0.02	0.86

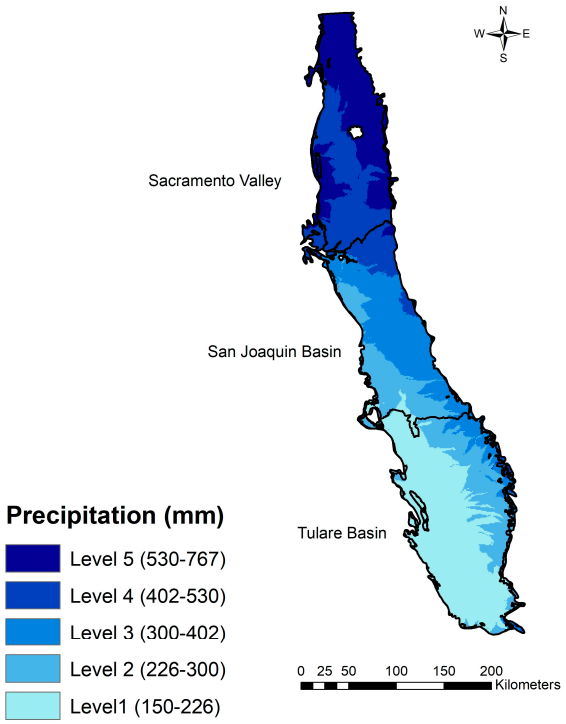

Figure 6. Distribution of rainfall in CV.

The average $PW_{N>5}$ at the low fertilizer level (Level 1) was 21.87% and showed an increasing trend of $PW_{N>5}$ towards the high fertilizer level through level 5 (55.97%). The difference between the mean $PW_{N>5}$ in Level 1 was significantly different from all the other higher fertilizer levels (Table S1 and Figure 7).

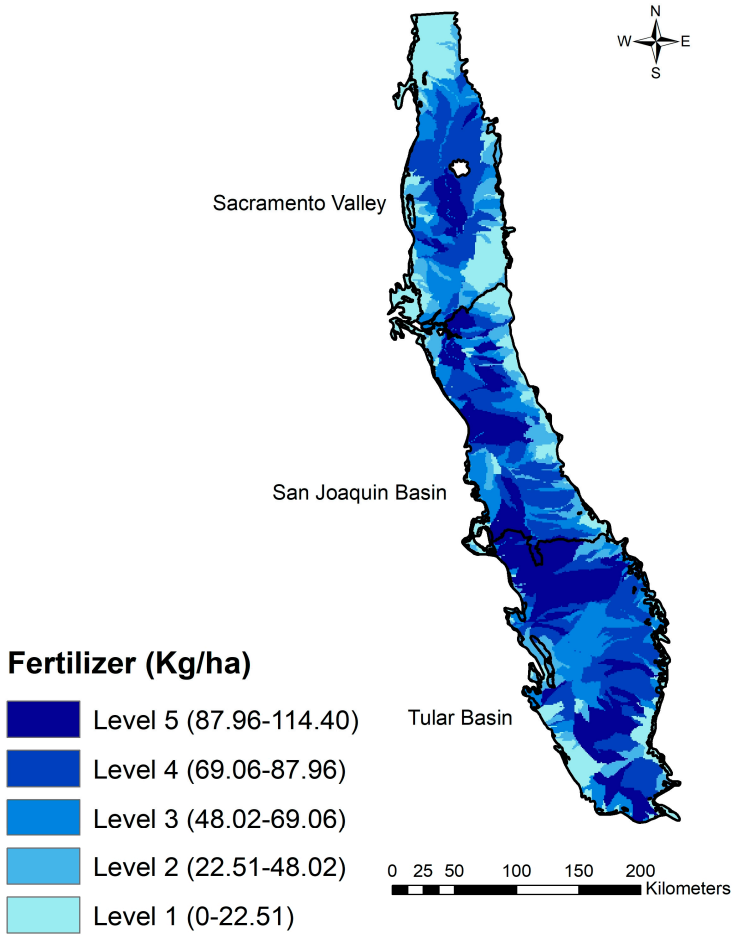

Figure 7. Distribution of fertilizer loading in the CV.

The average $PW_{N>5}$ in manure Level 1 was 29.46% and showed an increasing trend with $PW_{N>5}$ of 52.90% in Level 4. The average $PW_{N>5}$ in Level 1 was significantly different from all the remaining higher levels of manure (Table S1 and Figure 8). The average $PW_{N>5}$ was highest at the elevation Level 2 (54.08%), which was significantly different from both low-level elevation (Level 1) and high-level elevation (Level 3) (Table S1). Most of the cropland (Figure 9) in the CV was found in the Level 2 range of elevation (Figure 10) which is mostly Tulare Basin.

Groundwater nitrate contamination decreased with increasing percent clay in the soil, as clay particles inhibit the percolation of water into the aquifer. The average $PW_{N>5}$ was 48.06% at Level 1, which dropped to 26.97% at Level 3 at a statistically significant level. The average $PW_{N>5}$ in Level 4 and 5 was lower than Levels 1 and 2, but the difference was not statistically significant (Table S1 and Figure 11).

Interaction detector results (Table S2) showed that precipitation interacted strongly with permeability (0.46), elevation (0.41) and dissolved oxygen (0.43). The fertilizer interaction value increased for precipitation (0.38), permeability (0.36) and manure (0.35). The manure interaction value also increased for permeability (0.40), precipitation (0.39). The percent clay value increased for precipitation (0.36), manure (0.33) and elevation (0.34).

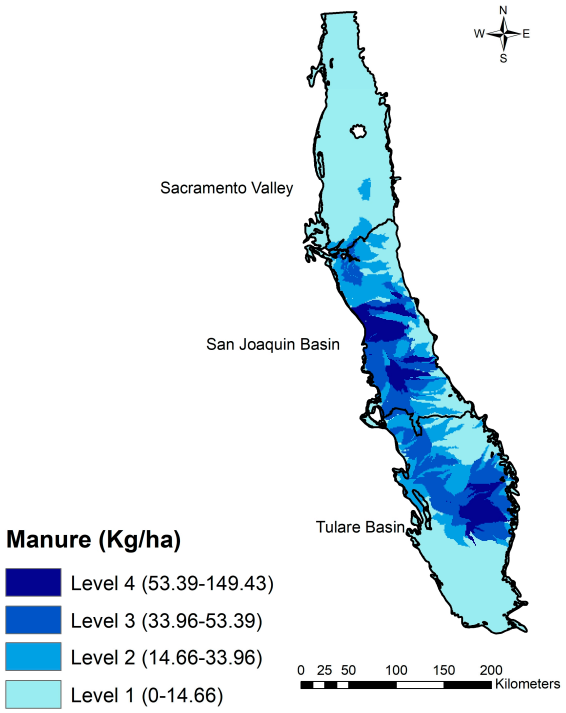

Figure 8. Distribution of manure loading in the CV.

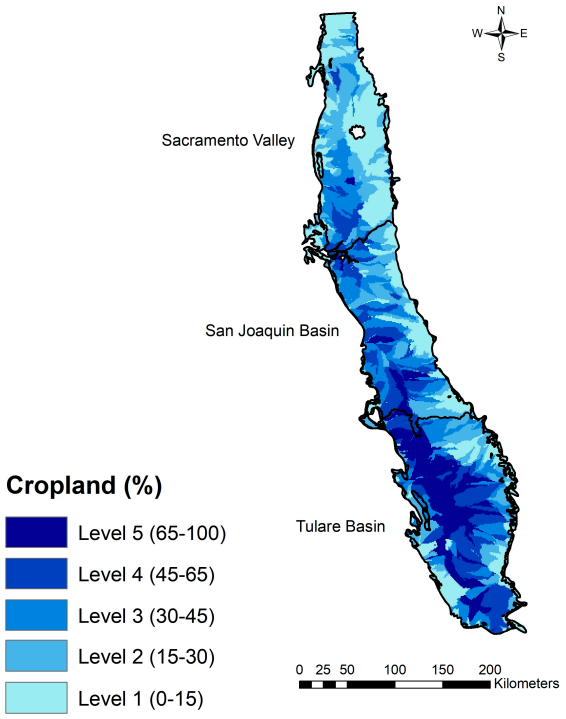

Figure 9. Percent of cropland in the CV.

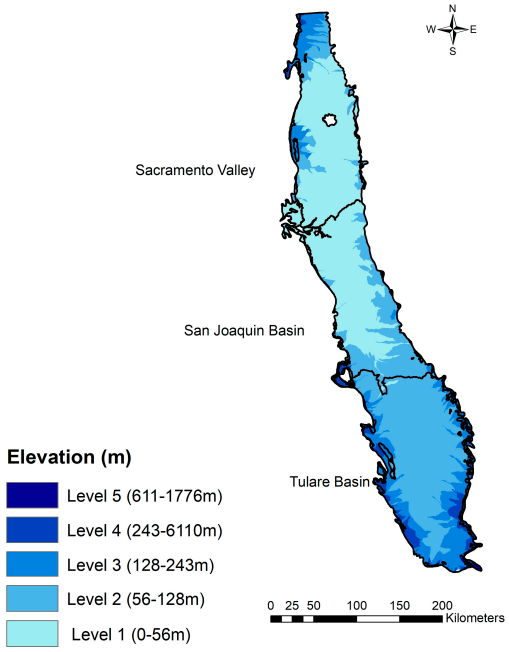

Figure 10. Elevation in the CV.

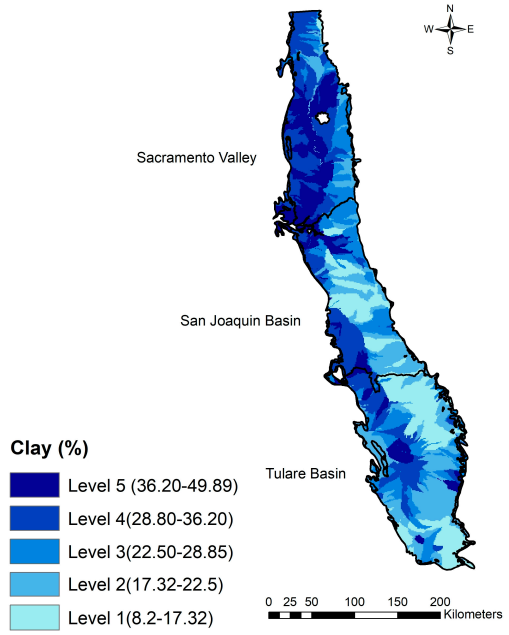

Figure 11. Percent of clay in the CV.

3.2. Principal Component Analysis

The Principal Component Analysis (PCA) results showed a Kaiser-Meyer-Olkin (KMO) value of 0.63 (Table 3). The score of 0.50 is suggested as the minimum KMO score to conduct PCA analysis. Therefore, our analysis meets the criteria of sampling adequacy. The Bartlett's test of sphericity was significant (<0.05), rejecting the null hypothesis that the correlation matrix is an identity matrix, meaning all the variables are uncorrelated. Correlation Analysis (Table 4) showed the highest correlation between cropland and manure (0.65), clay and permeability (−0.64), cropland and precipitation (−0.58), manure and fertilizer (0.50). It can be noted that fertilizer, manure and cropland are positively correlated with each other. This points to the fact that fertilizer and manure are applied in

the cropland and therefore the correlation between them is high. Precipitation is negatively correlated with fertilizer, manure and cropland, suggesting that rainfall is low in this area. The correlation between clay and permeability is high but negative. This is because the percolation of groundwater through a permeable area is high as water can easily percolate through its pore spaces, whereas clay inhibits the flow of water due to its small size particles. Iron and manganese are negatively correlated with dissolved oxygen because iron and manganese are found in low oxygen areas. Dissolved oxygen is negatively correlated with clay for the same reason, as clay size particles are small and have less pore space available for air.

Table 3. KMO and Bartlett's Test for PCA analysis.

Kaiser-Meyer-Olkin Measure of Sampling Adequacy.	0.61
Bartlett's Test of Sphericity Approx. Chi-Square	547.63
Df	66
Sig	0.00

Table 4. Correlation Matrix for explanatory variables.

	SLP	RECH	PPT	PERM	MN	MANU	FERT	FE	ELVN	DO	CROP	CLAY
SLP	1.00											
RECH	0.27 **	1.00										
PPT	0.21 *	0.11	1.00									
PERM	−0.11	0.19 *	−0.17	1.00								
MN	−0.15	−0.30 **	−0.13	−0.23 *	1.00							
MANU	−0.17	−0.14	−0.49 **	0.17	0.10	1.00						
FERT	−0.40 **	−0.27 **	−0.43 **	0.34 **	0.15	0.50 **	1.00					
FE	0.08	−0.19 *	0.38 **	−0.20 *	0.03	−0.17	−0.05	1.00				
ELVN	0.36 **	0.29 **	−0.43 **	−0.01	−0.07	−0.08	−0.08	−0.27 **	1.00			
DO	−0.10	0.55 **	−0.26 **	0.33 **	−0.15	0.08	0.05	−0.42 **	0.20 *	1.00		
CROP	−0.26 **	−0.31 **	−0.58 **	0.00	0.28 **	0.65 **	0.40 **	0.08	0.04	−0.09	1.00	
CLAY	0.05	−0.30 **	0.29 **	−0.64 **	0.45 **	−0.14	−0.16	0.30**	−0.26 **	−0.45 **	−0.01	1.00

** Correlation is significant at the 0.01 level. * Correlation is significant at the 0.05 level. (See Table 2 for full variable names).

Based on the scree plot and Eigen value greater than 1 (Table 5), only 4 components were retained. PC-1 explained 25.75%, PC-2 explained 23.21%, PC-3 explained 12.57% and PC-4 explained 8.96% of the variance. The cumulative percent of the variance explained by these four components was 70.50%. Table 6 shows variables significantly contributing to each PC, with loading >0.60. For PC-1, fertilizer, cropland and manure all have positive loadings, while precipitation has a negative loading.

Table 5. Total variance explained using the PCA method.

Total Variance Explained									
Component	**Initial Eigenvalues**			**Extraction Sums of Squared Loadings**			**Rotation Sums of Squared Loadings**		
	Total	**% of Variance**	**Cumulative %**	**Total**	**% of Variance**	**Cumulative %**	**Total**	**% of Variance**	**Cumulative %**
1	3.09	25.76	25.76	3.09	25.76	25.76	2.81	23.39	23.39
2	2.79	23.22	48.97	2.79	23.22	48.97	2.02	16.85	40.24
3	1.51	12.58	61.55	1.51	12.58	61.55	2.00	16.69	56.93
4	1.08	8.96	70.51	1.08	8.96	70.51	1.63	13.58	70.51
5	0.80	6.63	77.14						
6	0.68	5.70	82.84						
7	0.65	5.41	88.25						
8	0.48	4.03	92.28						
9	0.32	2.67	94.95						
10	0.25	2.11	97.06						
11	0.18	1.48	98.54						
12	0.18	1.46	100.00						

Table 6. Rotated Component Matrix.

	Component			
	1	**2**	**3**	**4**
CROP	0.85			
PPT	−0.80			
MANU	0.78			
FERT	0.67			
CLAY		−0.80		
PERM		0.76		
MN		−0.79		
DO			0.80	
FE			−0.79	
RECH				
ELVN				0.79
SLP				0.78

See Table 2 for full variable names.

Precipitation is low in San Joaquin Basin and decreases even more towards Tulare Basin. However, these are also the areas of high fertilizer, high manure and high cropland. Although rainfall is low, agriculture is performed here through irrigation by pumping groundwater from wells. Therefore, fertilizer, manure and cropland have positive loading and precipitation has a negative loading. This opposite relationship is also indicated by the correlation matrix where there is a negative correlation of precipitation with cropland (−0.58), fertilizer (−0.43) and manure (−0.49).

In PC-2, clay has the highest loading value of −0.80 followed by permeability (0.76) and manganese (−0.73). In areas of high permeability, nitrate dissolved in groundwater can easily move through the intergranular space between the rocks and can contaminate the aquifer. Therefore, permeability has high positive loadings. On the other hand, clay inhibits the flow of water as it is composed of very fine particles and hence it has a negative loading. Figure 12 shows that high permeable areas in the CV have low percent clay. Manganese, iron and dissolved oxygen represent the geochemical condition of the groundwater in the aquifer. Under anoxic conditions, nitrate can easily break down into nitrogen gas and reduce the concentration of nitrate in groundwater. This is because in the absence of oxygen, microbes in the groundwater preferentially use nitrate as an electron acceptor as a part of the redox reaction to produce energy. Therefore, the oxygen level represents the redox condition of the aquifer. Iron and manganese are also produced as a part of the redox reaction in groundwater, and their concentration increases under anoxic conditions. Therefore, manganese has a negative loading as it represents the anoxic condition, which can break down groundwater nitrate.

This is also exhibited by PC-3, where dissolved oxygen has a positive loading, as nitrate concentration in groundwater is higher if the dissolved oxygen is higher and is low under anoxic conditions (Figure 13). The nitrate concentration generally decreases towards the deeper aquifer as anoxic conditions develop and denitrification takes place [13]. PC- 4 shows elevation and slope both have positive loadings. Slopes are higher at higher elevation. The contamination of groundwater from nitrate is higher at lower elevation as the depth to the water table decreases at lower elevation.

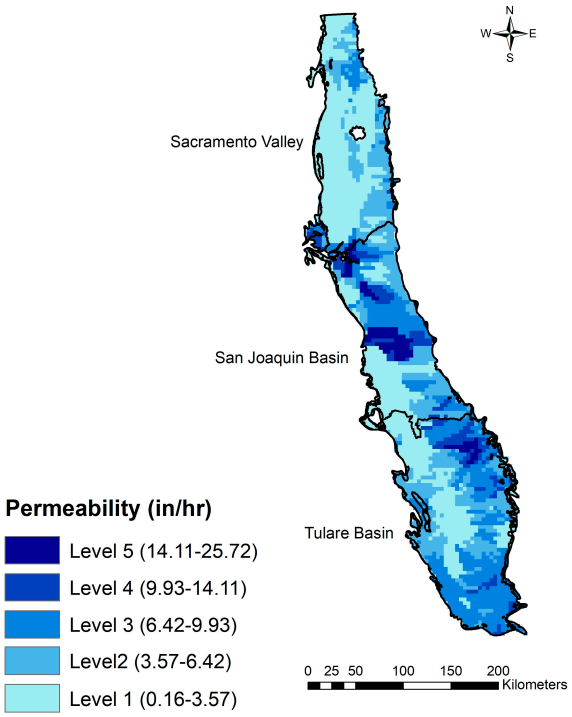

Figure 12. Distribution of permeability in the CV.

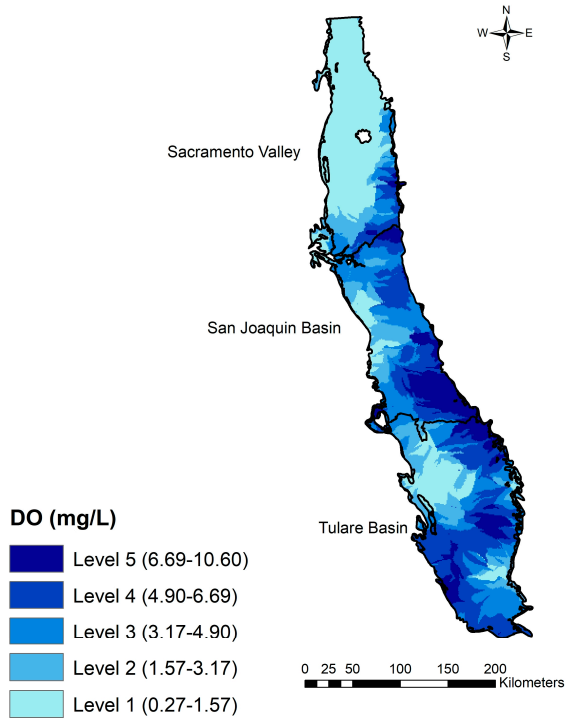

Figure 13. Distribution of dissolved oxygen in the CV.

3.3. Geographically Weighted Regression

The OLS method was performed prior to the GWR method and showed that the Variance Inflation Factor (VIF) of the variables was less than 7.5 with no multicollinearity. The Jarque-Bera test was not significant (p-value = 0.16), indicating that residuals are not normally distributed. A normally distributed residual means the model is biased and may be missing a key explanatory variable.

The adjusted-R^2 value is 0.21 and AIC value is 1611.28. R^2 measures the percent variance measured by the model, and the Akaike Information Criterion (AIC) measures the model performance. The spatial autocorrelation (Moran's I) value was 0.026 and was not significant, confirming residuals were not spatially clustered. Joint F-statistics were significant (p-value < 0.05), showing the relationship between the dependent variable and explanatory variables were non-stationary and could be improved by the GWR model. Precipitation, fertilizer and elevation were the significant variables (p-value < 0.05) in the OLS model. The coefficient for precipitation was negative, and the coefficient for manure and elevation was positive with $PW_{N>5}$.

In this GWR model, Adaptive Kernel was used to provide the geographic weighting in the model as the observations were not regularly positioned in the study area. The output of the GWR shows that the number of nearest neighbors that have been used in the estimation of each set of coefficients is 119 and estimates the percent of data under each kernel. The residual square value is 109,976, which is the sum of the square residual. A smaller residual square means that the GWR mode. fits well with the observed data. The effective number, which measures the complexity of the model, is 16.63. The Sigma value (16.63) is the estimated standard deviation for the residuals, and small values are preferred.

The Akaike Information Criterion (AIC) was used as the bandwidth method, as this method automatically finds the bandwidth, minimizes the AIC value, and gives the best predictions. The AIC value decreased from 1611.28 in the OLS model to 1606.96 in GWR, showing an increase in the fit of the model (Table 7).

Table 7. Comparison between OLS and the GWR model.

			Spatial Autocorrelation		
	Adj R^2	**AIC**	**Morans I**	**Z Score**	***p*-Value**
OLS	0.21	1611.28	0.026	0.49	0.62
GWR	0.26	1606.96	−0.00028	0.05	0.96

Generally, models with a lower AIC value and a decrease in the AIC value of more than 3 is considered better. Here, the AIC value of GWR dropped by 5 in comparison with OLS. The GWR model also showed improvement over the OLS model in the overall fit of the model by improving the adjusted-R^2 value from 0.21 to 0.26. However, the local R^2 values ranged from 0.024 to 0.314. Based on the local R^2 values, the model accurately predicted around Sacramento Valley and San Joaquin Basin, but local R^2 values were low around Tulare Basin.

The standardized residual map of GWR showed five watersheds which had standardized residuals greater than 2 (model under-prediction) and only one watershed with a negative standardized residual (model over-prediction). Spatial Autocorrelations (Moran's I) was computed to test if the residuals are random. The value of Moran's I ranges from −1 (complete dispersion) to +1 (complete clustering). Moran's I of value 0 indicates complete spatial randomness. Moran's I for GWR is −0.00028, with p-value 0.96, indicating that we cannot reject the null hypothesis that there is no spatial autocorrelation but accept that residuals are randomly distributed. Any spatial dependencies which might have been present in the OLS model were removed by the GWR model (Figures 14 and 15). The local condition (COND) number of the GWR model ranged from 24 to 27 throughout the CV. Since the COND number is less than 30, there is no local multicollinearity.

The global coefficient for precipitation is −5.6, indicating overall there is a negative relationship between the number of nitrate-contaminated wells and precipitation, perhaps due to the dilution effect of the precipitation in the CV. However, the local map of precipitation coefficients shows (Figure 16) that the value ranged from −13.63 in the north (Sacramento Valley) to 4.10 in the south (Tulare Basin), revealing the spatial heterogeneity of the model. Sacramento receives higher rainfall than the southern parts of the CV and therefore could experience a dilution effect of nitrate in groundwater. Tulare Basin and San Joaquin Basin receive low rainfall, and irrigation is performed by excessive pumping of

groundwater from wells. The depth to the water table is deeper in this part of the CV, and the groundwater flows towards the cone of depression. Also, the fertilizer and manure application rates are higher in this area. These conditions facilitate the percolation of nitrate into the aquifer. The local R^2 around this area also suggests that some other influential parameter could be added to the model.

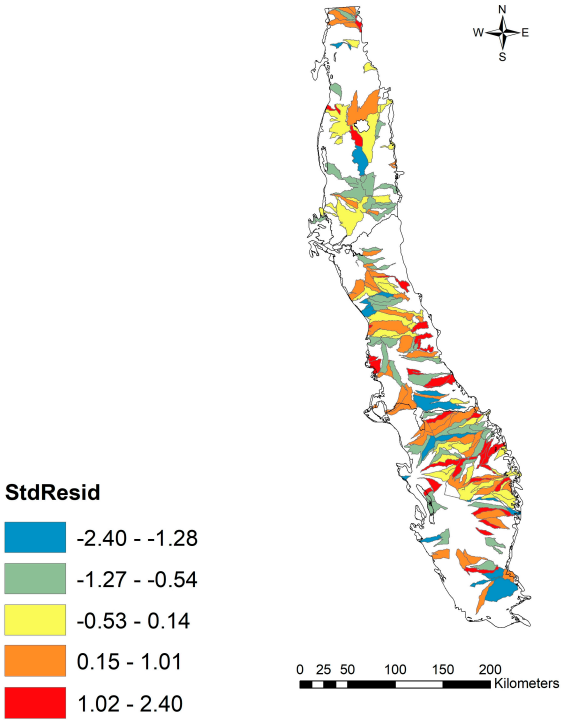

Figure 14. Standardized residuals of the OLS model.

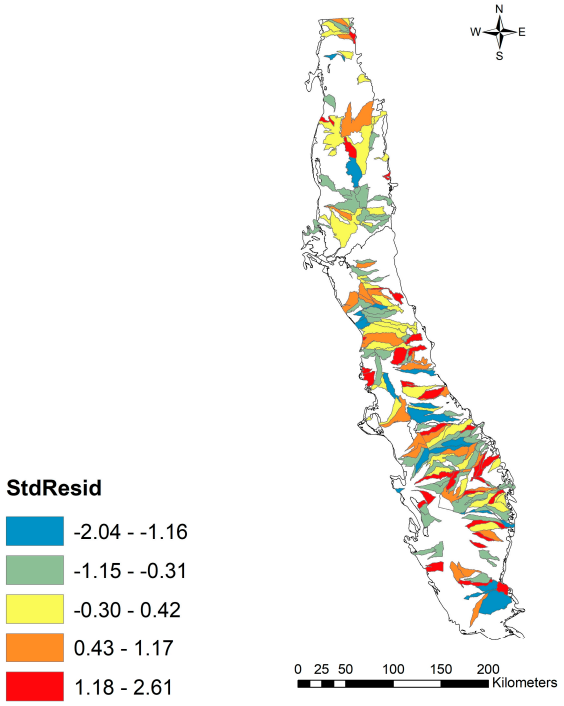

Figure 15. Standardized residuals of the GWR model.

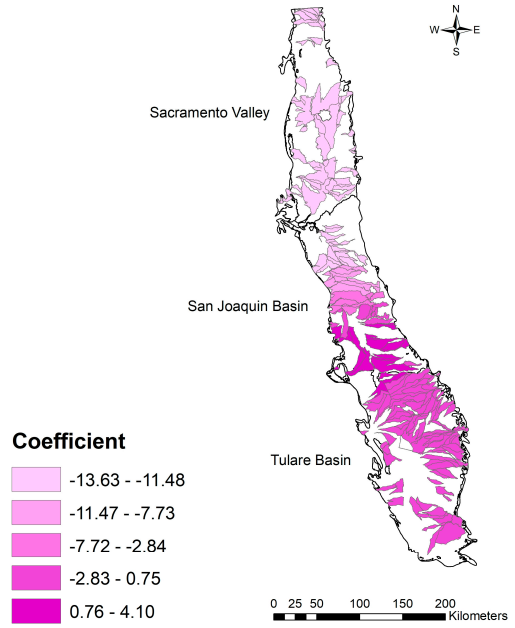

Figure 16. Coefficients of precipitation in the CV.

The global coefficient for the fertilizer is 4.05, indicating an overall positive relationship between nitrate contamination and fertilizer application. All three hydrologic regions in the CV have watersheds with higher fertilizer loadings. The local coefficient map (Figure 17) revealed some spatial heterogeneity in the model. The local coefficient varied from 0 to 6.7. Fertilizer coefficients were relatively higher around Tulare Basin, consistent with heavy irrigation and fertilizer application followed by Sacramento Valley and San Joaquin Basin.

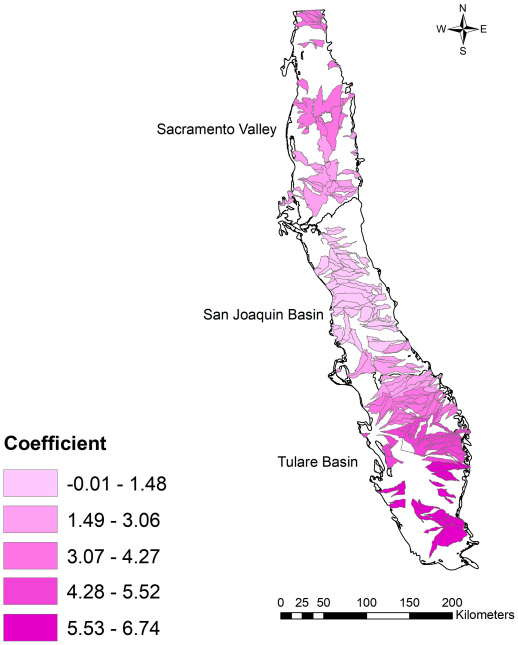

Figure 17. Coefficient of fertilizer in the CV.

The global coefficient for elevation (Figure 18) is 10.77. The local coefficient ranged from 0 to 22.4. The local coefficient is lowest around Tulare Basin and highest around San Joaquin Basin. Elevation is lowest around Sacramento Valley and San Joaquin Basin, which is more vulnerable to groundwater

contamination. This is consistent with the highest local coefficient values there. In general, the coefficient maps of GWR revealed the spatial heterogeneity of significant variables precipitation, fertilizer, and elevation, showing that local effects of groundwater contamination are prevalent in the CV.

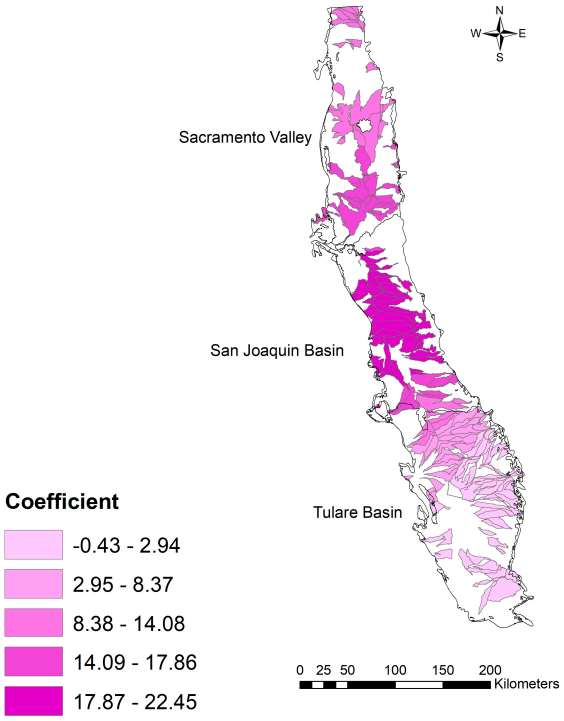

Figure 18. Coefficient of elevation in the CV.

4. Discussion

This study employed three different methods: GED, PCA and GWR to analyze the groundwater contamination of nitrate in the CV aquifer. The GED method identified precipitation, fertilizer, elevation, manure and clay as the significant variables contributing to the groundwater nitrate contamination. It showed geographic areas of high fertilizers, high manure, low percent clay and low rainfall areas (mainly San Joaquin Basin and Tulare Basin) as high average $PW_{N>5}$. The interaction detector indicated that precipitation interacted well with fertilizer and manure to enhance nitrate contamination as reflected by the higher combined PD value.

The PCA method was used to group the explanatory variables into four different components. The first four components with Eigenvalues greater than 1 explained 70.50% of the variances. The explanatory variables were grouped into different components, and their loading values in each component explained the contribution of each variable into the components. Fertilizer, manure, cropland and precipitation were grouped as redundant variables in PC-1. The positive and negative loadings on PC and the correlation matrix revealed the relationship between explanatory variables.

The GWR method captured the spatial heterogeneity of explanatory variables in the CV. This method improved the model compared to the OLS model by increasing the adjusted R-square and decreasing the AIC value. This method also reduced the spatial autocorrelation of residuals, strengthening the model. The coefficient map of precipitation, fertilizer and elevation was very informative in analyzing how the contribution of explanatory variables to $PW_{N>5}$ changes over the area of the CV.

The utilization of three different methods helped to demonstrate the groundwater nitrate contamination process in the CV from different perspectives. The application of three different methods not only allowed comparison between the methods but also extracted more information about the

contamination process in the CV. While GED identified key contributing explanatory variables and vulnerable geographic areas, PCA grouped the variables into different PCs to understand them in aggregation as they measure the same construction in the data. They are also known as redundant data and often cause multicollinearity in data analysis. For example, GED did not detect cropland as a significant variable, but PC-1 listed cropland, fertilizer, manure and precipitation as variables with loadings >0.60. From PC-1, it can be inferred that fertilizer and manure are applied in the cropland in areas with low rainfall, as precipitation has a negative loading. Also, the risk detector in the GED method showed increasing $PW_{N>5}$ as the level of cropland increased. Similarly, clay was significant in the GED method, and PC-2 also listed clay, permeability and manganese with high loading values. Clay (negative loading) and permeability are the reverse of each other as permeable layers have higher pore spaces and are often connected to facilitate groundwater flow, whereas clay consists of small sized particles that make water percolation difficult. Manganese also has a negative loading as this element is found under reducing conditions in high clay areas. Dissolved oxygen and iron were not significant in the GED method but had high loading in PC-3. Iron is found under reducing conditions (low oxygen); therefore, it has a negative loading compared to dissolved oxygen. $PW_{N>5}$ also progressively increased at a high dissolved oxygen level as nitrate is found under aerobic conditions, and the denitrification process breaks down nitrate into nitrogen gas under low dissolved oxygen. Elevation and slope both have positive loadings on PC-4 and elevation was also significant in the GED method.

The GWR, as a statistical method, captured the spatial heterogeneity of the explanatory variables and the overall fit of the model increased for the GWR model as compared to the OLS model. The spatial autocorrelation also showed residuals of the predictor variables that were more random in the GWR model. The OLS model did not show manure as a significant variable, but the GED method did. There is a significant difference in the $PW_{N>5}$ between the area of low manure and the area of high manure. The coefficient map of precipitation showed a slightly positive coefficient in the southern part of San Joaquin Basin, with a generally clear trend of a slightly negative coefficient in Tulare Basin and more negative coefficient in Sacramento Valley. Since the precipitation is low around Tulare Basin and increases towards northern Sacramento Valley, this could indicate that precipitation is diluting the groundwater nitrate contamination in the well samples. This trend was also exhibited by the GED model, showing decreasing $PW_{N>5}$ as the precipitation increases from south to north in the CV. The coefficient map of fertilizer also showed a higher coefficient in Tulare Basin and Sacramento Valley. There are numerous watersheds with high fertilizer rates in this part of the CV. The risk detector in the GED method also calculated an increasing average $PW_{N>5}$ at five different levels of increasing fertilizer rate. The coefficient map of elevation was high in Sacramento Valley and San Joaquin Basin. The San Joaquin Basin and Sacramento Valley had the highest positive coefficient, indicating its contribution to the groundwater nitrate. The GED method also revealed high $PW_{N>5}$ in Sacramento Valley and San Joaquin Basin.

The simultaneous observation of three different results provides strong confidence in the findings of these methods, either by complimenting each other or reconfirming the findings, which gives more confidence to the researcher about the contamination process in the CV. In addition, the findings of the interaction detector in the GED method point to the increased non-linear enhancement of precipitation with permeability (0.46), elevation (0.41), dissolved oxygen (0.43), fertilizer (0.38), manure (0.39) and clay (0.36), and confirms that Tulare Basin and San Joaquin Basin, as low rainfall areas, interact strongly with other variables in enhancing the groundwater nitrate contamination in the CV. However, careful analysis is required when analyzing data using different methods, as each method might have its own limitations. The GED method has a drawback of producing different results when the interval of explanatory variables changes, which defines a different geographical area. This study used natural break classification as the optimum classification method based on [57]. The PCA method also requires several assumptions like continuous variables, linear relationship between variables, sampling adequacy, correlation between variables and removal of significant outliers before the test is performed. Detail knowledge about the variables is also required to be able to explain the output,

as redundant variables are grouped into principal components. The correlation matrix in PCA aids in the understanding of relationships between explanatory variables. The GWR method also has the limitation of multicollinearity and kernel bandwidth selection. The regression residuals need to be analyzed carefully to make sure that there is no spatial autocorrelation to prevent model under- and over-prediction. The coefficient map should be carefully analyzed to observe its influence on dependent variables over the entire spatial region.

The results of this study are consistent with previous findings of groundwater nitrate contamination in the CV. The San Joaquin and Tulare Basin have been identified as regions vulnerable to groundwater nitrate contamination in the CV due to excess fertilizer and other hydrogeologic conditions. Recent studies applying machine learning methods such as boosted regression tree, artificial neural network and Bayesian network to nitrate data for shallow wells also predicted higher nitrate concentrations in Tulare Basin and San Joaquin Valley. Fertilizer, cropland and landuse data have been collected around sampled wells by establishing a buffer region around them in many studies [9,28]. Here, we analyzed the variables at the watershed level to closely represent the natural boundary and used $PW_{N>5}$ to understand the overall status of the watershed. This study provides a consistent outlook on the significant explanatory variables as well as how they vary over the CV, the interaction between the variables and data redundancy. These are added advantages of this study compared to previous studies, which used only non-parametric tests to compare average values of nitrate across different landuse types, performed temporal analysis comparing the decadal changes in nitrate concentration or analyzed the data above a threshold level. Statistical methods are widely used nowadays to analyze groundwater nitrate contamination data over a large spatial scale; however, it is still a challenge to find long-term and spatially well-covered data. Consistent temporal and spatial coverage of data is essential to any statistical analysis to remove bias in the study. Regular long-term monitoring of wells and data consistency are required to ensure effective statistical inference.

Overall, this comparative study showed that the GED method with its four different detectors was able to capture all the information revealed by the PCA and GWR methods. In addition, the GED method has the following advantages over the PCA and GWR methods. (1) It does not make assumptions about the data, in contrast to PCA or GWR, leading to easier data preparation and wider applicability; (2) It not only reveals which factors are more important but also which areas of each variable are more vulnerable; (3) It can examine the interaction between different factors; (4) It works well with both categorical and continuous data; (5) It is easier to interpret the contribution of individual variables and is thus more useful for policy makers.

5. Conclusions

This study is focused on the comparison of three different statistical methods: GED, PCA and GWR to study the groundwater nitrate contamination in the CV. The study was conducted at the watershed level in the CV by examining the relationship between nitrate contamination (as measured by percent of wells above a background concentration) in each watershed and twelve explanatory variables. The GED method was very effective in determining precipitation, fertilizer, elevation, manure and clay as the significant variables. A higher percent of contaminated wells was observed in areas where percent of cropland, fertilizer and manure were high. San Joaquin and Tulare Basin have a higher number of wells contaminated with nitrate even though annual precipitation is relatively low. This can be attributed to the higher fertilizer application rate in agricultural land and to changing hydrologic conditions over the years from groundwater pumping that could enhance the downward percolation of nitrate into the aquifer. The PCA results grouped precipitation, manure, fertilizer and cropland with high loadings on PC-1, suggesting their similar construct in the data set and that could cause multicollinearity problem in analysis. The GWR method captured the spatial heterogeneity of precipitation, fertilizer and elevation with respect to $PW_{N>5}$. The local coefficients of precipitation were only slightly positive in San Joaquin Basin and Tulare Basin, and an increasing negative coefficient was observed towards northern Sacramento Valley. The coefficients of fertilizer were high around Tulare

Basin where more fertilizer is applied. The coefficient of elevation was highest around Sacramento and San Joaquin Basin as these are relatively low elevation areas which are more susceptible to aquifer contamination.

Overall, the GED method was found to have added advantages over PCA and GWR as the results of GED are manifold compared to PCA and GWR. PCA is often used to reduce redundant data into fewer dimensions. However, PCA makes it difficult to interpret the grouped variables. The GWR model acts as a spatial microscope to reveal spatial changes of variables over the region but still suffers from multicollinearity that needs to be addressed using the OLS model. Although GWR can measure the changing coefficients over the study area, the GED method can directly compare different geographical areas based on the average value of the nitrate incidence rate. The less strict data requirement in the GED method in conjunction with its multiple outputs makes it an effective tool in understanding the overall nitrate contamination process in the Central Valley and beyond.

Supplementary Materials: The following are available online at www.mdpi.com/2220-9964/6/10/297/s1. Table S1: Risk Detector; Table S2: Interaction Detector; Table S3: Ecological Detector.

Acknowledgments: We thank the reviewers for their time spent reviewing the paper. We thank Xuwei Chen and Melissa Lenczewski for helpful discussions.

Author Contributions: Anil Shrestha and Wei Luo conceived and designed the research; Anil Shrestha performed data processing and analysis; Wei Luo contributed to the interpretation of the results; both authors wrote, revised, and improved the proposed article.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Reilly, T.E.; Dennehy, K.F.; Alley, W.M.; Cunningham, W.L. *Ground-Water Availability in the United States*; U.S. Geological Survey: Reston, VA, USA, 2008.
2. Faunt, C.C.; Belitz, K.; Hanson, R.T. *Groundwater Availability of the Central Valley Aquifer, California*; U.S. Geological Survey: Reston, VA, USA, 2009.
3. Bertoldi, G.L.; Johnston, R.H.; Evenson, K.D. *Ground Water in the Central Valley, California: A Summary Report*; United States Government Printing Office: Washington, WA, USA, 1991.
4. Rosentock, T.S.; Liptzin, D.; Dzurella, K.; Fryjoff-Hung, A.; Hollander, A.; Jensen, V.; King, A.; Kourakos, G.; McNally, A.; Pettygrove, G.S.; et al. Agriculture's Contribution to Nitrate Contamination of Californian Groundwater (1945–2005). *J. Environ. Qual.* **2014**, *43*, 895–907. [CrossRef] [PubMed]
5. Maupin, M.A.; Barber, N.L. *Estimated Withdrawals from Principal Aquifers in the United States, 2000*; U.S. Geological Survey: Reston, VA, USA, 2005.
6. Harter, T.; Onsoy, Y.S.; Heeren, K.; Michelle, D.; Gary, W.; Jan, W.H.; William, R.H. Deep vadose zone hydrology demonstrates fate of nitrate in eastern San Joaquin Valley. *Calif. Agric.* **2005**, *59*, 124–132. [CrossRef]
7. 100-Meter Resolution Color-Sliced Elevation of the Conterminous United States. Available online: https://catalog.data.gov/dataset/100-meter-resolution-color-sliced-elevation-of-the-conterminous-united-states-direct-download (accessed on 25 September 2017).
8. Dubrovsky, N.M.; Kratzer, C.R.; Brown, L.R.; Gronberg, J.M.; Burrow, K.R. *Water Quality in the San Joaquin-Tulare Basins, California 1992–1995*; U.S. Geological Survey Cirular 1159; U.S. Geological Survey: Denver, CO, USA, 1998.
9. Lockhart, K.; King, A.M.; Harter, T. Identifying sources of groundwater nitrate contamination in a 2 large alluvial groundwater basin with highly diversified 3 intensive agricultural production. *J. Contam. Hydrol.* **2013**, *151C*, 140–154. [CrossRef] [PubMed]
10. Spalding, R.F.; Exner, M.E. Occurrence of Nitrate in Groundwater—A Review. *J. Environ. Qual.* **1993**, *22*, 392–402. [CrossRef]
11. Dubrovsky, N.M.; Burrow, K.R.; Clark, G.M.; Gronberg, J.M.; Hamilton, P.A.; Hitt, K.J.; Mueller, D.K.; Munn, M.D.; Nolan, B.T.; Puckett, L.J.; et al. *The Quality of Our Nation's Waters—Nutrients in the Nation's Streams and Groundwater, 1992–2004*; U.S. Geological Survey: Reston, VA, USA, 2010.

12. Burrow, K.R.; Nolan, B.T.; Rupert, M.G.; Dubrvosky, N.M. Nitrate in Groundwater of the United States, 1991–2003. *Environ. Sci. Technol.* **2010**, *44*, 4988–4997. [CrossRef] [PubMed]
13. Burrow, K.R.; Jurgens, B.C.; Belitz, K.; Dubrovsky, N.M. Assessment of regional change in nitrate concentrations in groundwater in the Central Valley, California, USA, 1950s–2000s. *Environ. Earth Sci.* **2012**, *69*, 2609–2621. [CrossRef]
14. Schans van der, M.; Harter, T.; Leijnse, A.; Mathews, M.C.; Meyer, R.D. Characterizing sources of nitrate leaching from an irrigated dairy farm in Merced County, California. *J. Contam. Hydrol.* **2009**, *110*, 9–21. [CrossRef] [PubMed]
15. Wright, M.T.; Belitz, K.; Johnson, T. *Assessing the Susceptibility to Contamination of Two Aquifer Systems Used for Public Water Supply in the Modesto and Fresno Metropolitan Areas, California, 2001 and 2002*; U.S. Geological Survey: Reston, VA, USA, 2004; p. 35.
16. Troiano, J.; Garretson, C.; Dasilva, A.; Marade, J.; Barry, T. Pesticide and Nitrate Trends in Domestic Wells where Pesticide Use Is Regulated in Fresno and Tulare Counties, California. *J. Environ. Qual.* **2013**, *42*, 1711–1723. [CrossRef] [PubMed]
17. Jurgens, B.C.; Burrow, K.R.; Dalgish, B.A.; Shelton, J.L. *Hydrogeology, Water Chemistry, and Factors Affecting the Transport of Contaminants in the Zone of Contribution of a Public-Supply Well in Modesto, Eastern San Joaquin Valley, California*; U.S. Geological Survey: Reston, VA, USA, 2008.
18. Harter, T.; Lund, J.R. *Addressing Nitrate in California's Drinking Water: With a Foucs on Tulare Lake Basin and Salinas Valley Groundwater*; Center for Watershed Sciences, University of California: Davis, CA, USA, 2012; p. 78.
19. Gardner, K.; Vogel, R.M. Predicting ground water nitrate concentration from land use. *Groundwater* **2005**, *43*, 343–352. [CrossRef] [PubMed]
20. Baker, R.J.; Chepiga, M.M.; Cauller, S.J. *Median Nitrate Concentrations in Groundwater in the New Jersey Highlands Region Estimated Using Regression Models and Land-Surface Characteristics*; U.S. Geological Survey Scientific Investigation Report; U.S. Geological Survey: Reston, VA, USA, 2015; p. 27.
21. Zhang, W.; Li, H.; Sun, D.; Zhou, L. A Statistical Assessment of the Impact of Agricultural Land Use Intensity on Regional Surface Water Quality at Multiple Scales. *Int. J. Res. Public Health* **2012**, *9*, 4170–4186. [CrossRef] [PubMed]
22. Nolan, B.T.; Hitt, K.J.; Ruddy, B.C. Probability of Nitrate Contamination of Recently Recharged Groundwaters in the Conterminous United States. *Environ. Sci. Technol.* **2002**, *36*, 2138–2145. [CrossRef] [PubMed]
23. Nolan, B.T.; Gronberg, J.M.; Faunt, C.C.; Eberts, S.M., II; Belitz, K. Modeling Nitrate at Domestic and Public-Supply Well Depths in the Central Valley, California. *Environ. Sci. Technol.* **2014**, *48*, 5643–5651. [CrossRef] [PubMed]
24. Fram, M.S.; Belitz, K. Probability of Detecting Perchlorate under Natural Conditions in Deep Groundwater in California and the Southwestern United States. *Environ. Sci. Technol.* **2011**, *45*, 1271–1277. [CrossRef] [PubMed]
25. Burrow, K.R.; Shelton, J.L.; Dubrovsky, N.M. *Occurance of Nitrate and Pesticide in Groundwater beneath Three Agricultural Land-Use Setting in the Eastern San Joaquin Valley, California. 1993–1995*; U.S. Geological Survey: Sacramento, CA, USA, 1998.
26. Burrow, K.R.; Shelton, J.L.; Dubrvosky, N.M. Regional nitrate and pesticide trends in ground water in the eastern San Joaquin Valley, California. *J. Environ. Qual.* **2008**, *37*, S249–S263. [CrossRef] [PubMed]
27. Ransom, K.M.; Nolan, B.T.; Traum, J.A.; Faunt, C.C.; Bell, A.M.; Gronberg, J.A.M.; Wheeler, D.; Rosecrans, C.Z.; Jurgens, B.C.; Schwarz, G.E.; et al. A hybrid machine learning model to predict and visualize nitrate concentration throughout the Central Valley aquifer, California, USA. *Sci. Total Environ.* **2017**, *601–602*, 1160–1172. [CrossRef] [PubMed]
28. Nolan, B.T.; Fienen, M.N.; Lorenz, D.L. A statistical learning framework for groundwater nitrate models of the Central Valley, California, USA. *J. Hydrol.* **2015**, *531*, 902–911. [CrossRef]
29. Duan, W.; He, B.; Nover, D.; Yang, G.; Chen, W.; Meng, H.; Zou, S.; Liu, C. Water Quality Assessment and Pollution Source Identification of the Eastern Poyang Lake Basin Using Multivariate Statistical Methods. *Sustainability* **2016**, *8*, 133. [CrossRef]
30. Kim, H.; Park, S. Hydrogeochemical Characteristics of Groundwater Highly Polluted with Nitrate in an Agricultural Area of Hongseong, Korea. *Water* **2016**, *8*, 345. [CrossRef]

31. Wu, T.-N.; Su, C.-S. Application of Principal Component Analysis and Clustering to Spatial Allocation of Groundwater Contamination. In Proceedings of the Fourth International Conference on Fuzzy Systems and Knowledge Discovery, Jinan, China, 18–20 October 2008; Volume 4.
32. Ul-Saufie, A.Z.; Yahya, A.S.; Ramli, N.A. Improving multiple linear regression model using principal component analysis for predicting PM10 concentration in Seberang Prai, Pulau Pinang. *Int. J. Environ. Sci.* **2011**, *2*, 403–408.
33. Camdevyren, H.; Demyr, N.; Kanik, A.; Syddyk, K. Use of principal component scores in multiple linear regression models for prediction of *Chlorophyll-a* in reservoirs. *Ecol. Model.* **2005**, *181*, 581–589. [CrossRef]
34. Ling, T.-Y.; Soo, C.-L.; Liew, J.-J.; Nyanti, L.; Sim, S.-F.; Grinang, J. Application of Multivariate Statistical Analysis in Evaluation of Surface River Water Quality of a Tropical River. *J. Chem.* **2017**, *2017*, 5737452. [CrossRef]
35. Amano, H.; Kei, N.; Ronny, B. Groundwater geochemistry of a nitrate-contaminated site. *Environ. Earth Sci.* **2016**, *75*, 1145. [CrossRef]
36. Fotheringham, A.S.; Brunsdon, C.; Charlton, M.E. *Geographically Weighted Regression: The Analysis of Spatially Varying Relationships*; Wiley: New York, NY, USA, 2003; ISBN 978-0-471-49616-2.
37. Tu, J. Spatially varying relationships between land use and water quality across an urbanization gradient explored by geographically weighted regression. *Appl. Geogr.* **2011**, *31*, 376–392. [CrossRef]
38. Pratt, B.; Chang, H. Effects of Land Cover, Topography, and Built Structure on Seasonal Water Quality at Multiple Spatial Scales. *J. Hazard. Mater.* **2012**, *209–210*, 48–58. [CrossRef] [PubMed]
39. Chang, H.; Psaris, M. Local landscape predictors of maximum stream temperature and thermal sensitivity in the Columbia River basin, USA. *Sci. Total Environ.* **2013**, *461–462*, 587–600. [CrossRef] [PubMed]
40. Javi, S.T.; Malekmohammadi, B.; Mokhtari, H. Application of geographically weighted regression model to analysis of spatiotemporal varying relationships between groundwater quantity and land use changes (case study: Khanmirza Plain, Iran). *Environ. Monit. Assess.* **2014**, *186*, 3123–3138. [CrossRef] [PubMed]
41. Wheeler, D.; Tiefelsdorf, M. Multicollinearity and correlation among local regression coefficients in geographically weighted regression. *J. Geogr. Syst.* **2005**, *7*, 161–187. [CrossRef]
42. Wang, J.-F.; Li, X.-H.; Christakos, G.; Liao, Y.-L.; Zhang, T.; Gu, X.; Zheng, X.-Y. Geographical Detectors-Based Health Risk Assessment and its Application in the Neural Tube Defects Study of the Heshun Region, China. *J. Int. J. Geogr. Inf. Sci.* **2010**, *24*, 107–127. [CrossRef]
43. Wang, J.-F.; Zhang, T.-L.; Fu, B.-J. A measure of spatila stratified heterogenity. *Ecol. Indic.* **2016**, *67*, 250–256. [CrossRef]
44. Luo, W.; Jasiewicz, J.; Stepinski, T.; Cang, X. Spatial association between dissection density and environmental factors over the entire conterminous United States. *Geophys. Res. Lett.* **2016**, *43*, 691–700. [CrossRef]
45. Shrestha, A.; Luo, W. An assessment of groundwater contamination in Central Valley aquifer, California using geodetector method. *Ann. GIS* **2017**, *23*, 149–166. [CrossRef]
46. Gronberg, J.A.M.; Saphr, N.E. *County-Level Estimates of Nitrogen and Phosphorus from Commercial Fertilizer for the Conterminous United States, 1987–2006*; US Geological Survey: Reston, VA, USA, 2012; p. 20.
47. Mueller, D.K.; Gronberg, J.A.M. *County-Level Estimates of Nitrogen and Phosphorus from Animal Manure for the Conterminous United States, 2002*; U.S. Geological Survey: Reston, VA, USA, 2013.
48. Homer, C.; Dewitz, J.; Fry, J.; Hossain, N.; Larson, C.; Coan, M.; Herold, N.; McKerrow, A.; VanDriel, J.N.; Eickham, J. National Landcover Database 2001 (NLCD 2001). Available online: https://www.mrlc.gov/nlcd2001.php (accessed on 12 August 2013).
49. Wolock, D.M. *STATSGO Soil Characteristics for the Conterminous United States*; U.S. Geological Survey: Reston, VA, USA, 1997.
50. PRISM Climate Data. Available online: http://www.prism.oregonstate.edu/ (accessed on 5 January 2013).
51. Elevation Derivatives for National Applications (EDNA). Available online: https://lta.cr.usgs.gov/edna (accessed on 10 January 2014).
52. Nolan, B.T.; Hitt, K.J. Vulnerability of shallow ground water and drinking-water wells to nitrate in the United States: Model of predicted nitrate concentration in shallow, recently recharged ground water—Input data set for clay sediment (gwava-s clay). *Environ. Sci. Technol.* **2006**, *40*, 7834–7840. [CrossRef] [PubMed]
53. Wolock, D.M. *Estimated Mean Annual Natural Ground-Water Recharge in the Conterminous United States*; U.S. Geological Survey Open-File Report; U.S. Geological Survey: Reston, VA, USA, 2003.

54. National Water-Quality Assessment (NAWQA) Project. Available online: https://water.usgs.gov/nawqa/ (accessed on 25 August 2013).
55. National Water Information System. Available online: https://waterdata.usgs.gov/nwis (accessed on 12 March 2014).
56. Nolan, B.T.; Hitt, K.J. *Nutrients in Shallow Ground Waters beneath Relatively Undeveloped Areas in the Conterminous United States*; National Water-Quality Assessment Program; U.S. Geological Survey: Denver, CO, USA, 2003.
57. Cao, F.; Ge, Y.; Wang, J.-F. Optimal discretization for geographical detectors-based risk assessment. *GIScience Remote Sens.* **2013**, *50*, 78–92.
58. Fotheringham, A.S.; Charlton, M.E.; Brunsdon, C. Spatial Variations in School Performance: A Local Analysis Using Geographically Weighted Regression. *Geogr. Environ. Model.* **2001**, *5*, 43–66. [CrossRef]