SPATIO-TEMPORAL ANALYSIS OF SURFACE WATER QUALITY: A CALIFORNIA CASE STUDY

Houlin Chen [1], Meredith Franklin [2,3]

[1] Faculty of Arts and Science, University of Toronto, Ontario, CA
[2] Department of Statistical Sciences and School of the Environment, University of Toronto, Ontario, CA
[3] Department of Population and Public Health Sciences, University of Southern California, CA, USA

ABSTRACT

Surface water quality has a direct impact on public health, ecosystems, and agriculture, in addition to being an important indicator of the overall health of the environment. This study offers a comprehensive assessment of these patterns by leveraging around 70 years of data in California, taking into account climate zones and geographical types. We analyzed surface water quality indicators, including pH, dissolved oxygen, specific conductance, and water temperature, based on field results from 5,080 water quality stations in California Water Quality Data (CWQD). Machine learning (ML) models were developed to establish relationships between spatial and temporal variables, climate zones, geographical types, and water quality indicators. Applying these models to spatially interpolate the four water quality indicators over California, the research results indicate an uneven distribution of water quality indicators in California, suggesting the presence of potential pollution zones, seawater erosion, and effects of climate change.

***Index Terms*—** Water quality, spatio-temporal modeling, climate, inland waters, water ecosystem

1. INTRODUCTION

The quality of surface water is an integral factor in various aspects of human and ecosystem life. It is affected by a diverse range of elements, from microbial content [1] to industrial effluents [2], to Earth's water cycle [3], all playing a significant role in determining the quality of surface water. A notable illustration of a significant water quality issue is the nuclear wastewater leak in Fukushima, Japan, in 2011 [4]. The effects of the leak reached the United States coastline in just three years due to atmospheric pressure and ocean currents, with repercussions on water quality that are projected to extend for over 30 years [5]. Water plays a crucial role in human well-being and health, making the investigation and prediction of water quality trends an essential area of research [6].

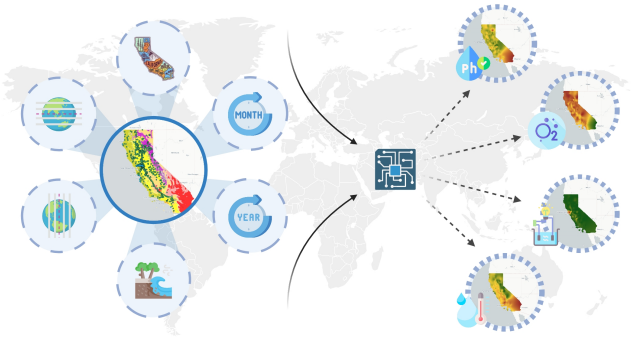


Fig. 1: System diagram for the spatio-temporal analysis of surface water quality.

California, a highly populated state characterized by its diverse climates and micro-climates, as well as varied geographical features, has been the subject of significant environmental research. It hosts three major climate types - arid, temperate, and cold [7], and also experiences seawater erosion on its coastal freshwater system. Water quality concerns in California have predominantly been about nitrate contamination from agricultural activities in the Central Valley such as fertilizer application [8]. Elevated concentrations of nitrate have been associated with adverse birth outcomes [9] and thyroid cancer [10]. Other markers of water quality, including those examined in this study, are important to monitor to ensure California's water is safe for human consumption and as a habitat for aquatic life, which indirectly affects health through the food chain. In a United States Geological Survey (USGS) report, pH measured at 1,337 wells over a 20-year period from 1993 to 2014 in California's Central Valley was modelled and mapped to provide an understanding of water quality conditions at domestic and public supply drinking water zones [11]. Another USGS report analyzed specific conductance and water temperature at eight stations in the San Francisco Bay Area from 1990 to 2015, found that both parameters reached record highs in the region in 2015 [12]. However, the existing literature on water quality in California lacks comprehensive spatio-temporal modeling that accounts for the long-term assessment and integration of multiple indi-

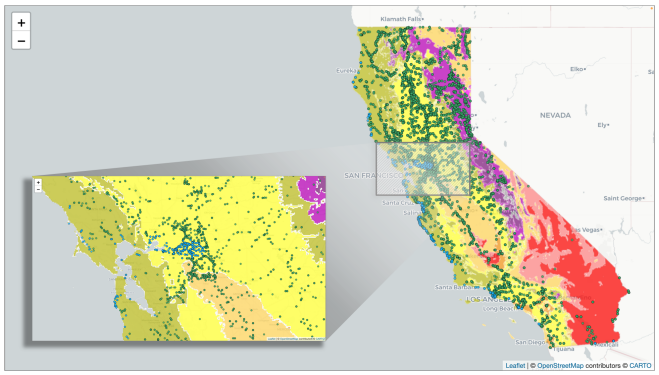


Fig. 2: Distribution of water quality stations in California. Points on the map represent stations, with Inland and Coastal in Geographical Type, respectively. The varied background colors delineate different climate zones based on the Köppen climate classification.

cators, and often overlooks the diverse range of climate zones and geographical types [13, 14].

In this paper, we conducted a spatio-temporal analysis of water quality indicators (pH, dissolved oxygen, specific conductance, and water temperature) collected in California over the past 70 years, spanning from 1956 to 2023. We establish relationships between spatio-temporal variables and these indicators using regression and machine learning (ML) models (Fig. 1). Through this approach we reveal potential interactions among water quality indicators, climate zones, and geographical types. The trained spatio-temporal models were applied to spatially interpolate the water quality indicators over the state, allowing us to visualize how ecological and climate changes in California have impacted surface water, which can suggest future environmental protection policies. For the complete version of this paper, please refer to [15].

2. MATERIALS AND METHODS

2.1. California Water Quality Data

The California Water Quality Dataset (CWQD) was sourced from the California Department of Water Resources [16]. It encapsulates a comprehensive collection of both field and laboratory results, underpinned by various physical and chemical parameters. Throughout its lifecourse, the CWQD has included a total of 29,229 water quality monitoring stations across the state (Fig. 2). This expansive scope covers diverse terrains, extending from deserts to islands and highlands to basins. The CWQD provides three distinct classifications of water based on varying depths: surface water, under-surface water, and groundwater. In this paper we focus on the 5,080 surface water quality monitoring stations, spanning from January 1956 to July 2023, with measured parameters including pH, Dissolved Oxygen, Specific Conductance, and Water Temperature. Most of the 5,080 stations have data on samples collected four times per month. Between 1960 and 2020, data collected by long-term existing water quality stations at 20-year intervals showed no regular trend in pH levels, an upward trend in water temperature, and a downward trend in the other two water quality indicators (Table 1).

Table 1: The 20-year average values of four water quality indicators for long-term existing water quality stations from 1960 to 2020.

	pH (pH Units)	**Dissolved Oxygen** (mg/L)	**Specific Conductance** ($\mu S/\text{cm@25}^\circ\text{C}$)	**Water Temperature** ($^\circ$C)
1960-1980	8.203	9.491	542.274	17.195
1980-2000	7.978	9.299	463.364	17.273
2000-2020	8.068	8.630	444.488	17.701

2.2. Köppen Climate and Geographical Types

Climate zones are inextricably linked to water quality due to the pronounced impact different temperatures and precipitation levels exert on water quality indices [17]. For example, varying temperatures can influence the solubility of substances in water, while differing precipitation levels may affect the concentration of dissolved substances by altering water flow and levels. We employed the Köppen climate classification [18] to attribute climate labels to each station in California (Fig. 2). The Köppen climate classification is a globally acknowledged and utilized climate categorization system, extensively applied in various disciplines such as meteorology, climatology, and environmental science, to facilitate tasks like regional climate categorization, biodiversity studies, and climatic impact analyses. The Köppen climate classification consists of five primary climate types: (A), arid (B), temperate (C), cold (D), and polar (E). Each of these major classifications is further refined into a total of 30 sub-climates based on nuanced differences in precipitation, temperature variations, and geographical features. The sub-climates are differentiated by factors such as the type of seasonal precipitation and the degree of heat, exemplified by classifications such as BSk, which stands for arid (B), steppe (S), and cold (k), and Csb, which represents temperate (C), dry summer (s), and warm summer (b). When formulating these variables for our models, we used both the broader classifications, incorporating only the pertinent three, as the climatic specifics of California do not involve tropical (A) or polar (E) classifications, as well as the finely segmented sub-climates, of which there were nine in across the state.

The composition and concentration of constituents vary significantly between seawater and inland waters, mainly due to the high salinity and pollutants found in seawater. Seawater intrusion in coastal areas impacts the specific conductance of inland surface water, which was of concern for our assessment of inland water quality properties. Since there is no conclusive agreement on the distance, intensity, and depth of seawater intrusion into inland surface waters in the liter-

Table 2: Comparison of various regression models for predicting water quality indicators on the test set. The metrics are RMSE (↓ lower is better) and R^2 (↑ higher is better). The best performance for each indicator is highlighted in **bold**.

	pH (pH Units)		**Dissolved Oxygen** (mg/L)		**Specific Conductance** ($\mu S/\text{cm@25}^\circ\text{C}$)		**Water Temperature** ($^\circ$C)	
	RMSE (↓)	R^2(↑)	RMSE (↓)	R^2(↑)	RMSE (↓)	R^2(↑)	RMSE (↓)	R^2(↑)
LM	0.498	0.207	1.913	0.230	405.445	0.210	4.516	0.339
RF	0.378	0.542	1.452	0.557	257.467	0.682	1.859	0.900
GP	0.465	0.309	1.856	0.275	346.392	0.425	2.757	0.757
SVM	0.428	0.415	1.649	0.428	380.100	0.306	2.221	0.842
GAM	0.432	0.402	1.715	0.381	348.542	0.417	2.306	0.830
XGBoost	**0.376**	**0.548**	**1.380**	**0.599**	**247.900**	**0.705**	**1.738**	**0.903**

ature [19, 20, 21], we propose categorizing stations located within eight kilometers of the coastline as Coastal stations, and those situated more than eight kilometers away as Inland stations (Fig. 2).

2.3. Data Pre-processing and Regression Model

In the data pre-processing phase, we addressed issues such as implausible non-negative longitude values and duplicate records with incorrect station IDs by setting their observations to missing. For the four water quality indicators we removed outliers by discarding data points that fell outside the 95% th percentile. After conducting these steps, we had a total of 64,185 samples, which we then partitioned into training and testing subsets at an 80:20% ratio, using the training set to develop the model and the unseen test set to evaluate its accuracy.

To predict each of the surface water quality indicators (i.e., pH, Dissolved Oxygen, Specific Conductance, and Water Temperature), we used a predictor variable set that included the Köppen climate zones, geographical types, latitude, longitude, year, month, and the other three water quality indicators. We focused on six regression models, namely the Linear Model (LM), Generalized Additive Model (GAM), Random Forest (RF), Gaussian Process (GP), Support Vector Machine (SVM) and Extreme Gradient Boosting (XGBoost). While LM and GAM represent more traditional statistical approaches to spatio-temporal modeling, machine learning has been recognized for its capacity to perform nonlinear parameter fitting and has been widely adopted in the estimation of spatio-temporal variables to discern complex patterns and interactions in data [22, 23, 24]. For the GAM model, which can be explicitly used for spatial and temporal interpolation, we used spatiotemoporal smoothing of latitude and longitude and time through a tensor product basis, as well as cubic regression splines to capture temporal trends [25].

To evaluate model performance, we employed the root mean square error (RMSE) and the coefficient of determination (R^2) as our principal metrics on the test set. RMSE provides insights into the discrepancies between the actual values in the dataset and the predictions from the model, while R^2 quantifies the explanatory power of the model.

3. RESULTS

3.1. Model Prediction

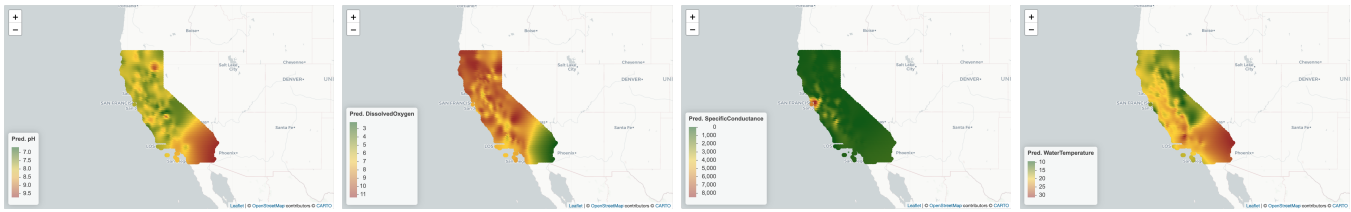
The performance of the six models evaluated using RMSE and R^2 (Table 2) show that XGBoost had the best performance across all four water quality indicators. Water temperature has the highest R^2 = 0.903, whereas it was lowest for pH R^2 = 0.548, indicating that, comparatively, pH predictions have lower accuracy and interpretability. Possibly due to errors in manual detection and data recording as well as the interference caused by external environmental factors, a significant error margin exists for pH. However, the data preprocessing we undertook, coupled with the model's inherent robustness against noise, we are confident that the model results fall within an acceptable range. Although not shown here, by incorporating the other water quality indicators and variables for climate zone, and geographical type significantly enhances the predictive accuracy of the models.

3.2. Spatial Interpolation

We use the GAM model to estimate the spatial distribution of water quality indicators in California for July 2023 as an example and demonstrate distribution maps of water quality indicators (Fig. 3). These indicators vary spatially but exhibit some intrinsic connections. For example, pH, Dissolved Oxygen, and Water Temperature all demonstrate a gradient variation from north to south, with specific characteristics in the desert region of southern California.

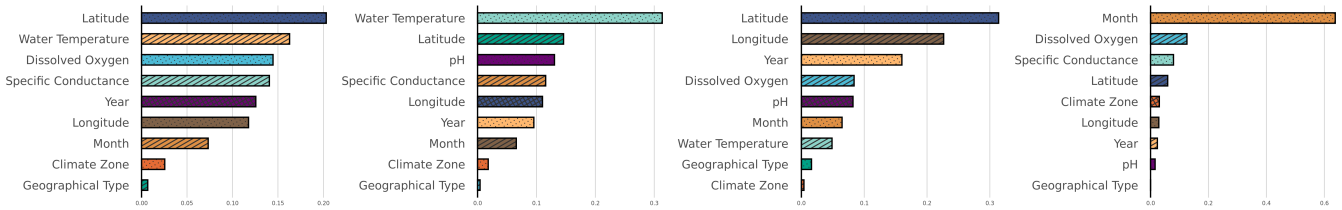
The pH values of surface water are slightly alkaline. In the southern California desert, pH levels are uniformly around 9, while other regions display variable pH distributions due to factors like temperature, geographical influences, and human activities (Fig. 3a). Varied pH values in the north and central parts are likely influenced by aquatic biological activity and agricultural irrigation. The elevated pH in the southern desert is possibly due to evaporation and soil salinity.

Dissolved Oxygen shows a gradient similar to pH, influenced by temperature and geographical features (Fig. 3b). Cooler waters typically have higher Dissolved Oxygen levels, and dense vegetation in mountainous areas contributes to this. However, some Central Valley regions exhibit lower levels, potentially due to ecological factors and human impact. Generally, most areas have high Dissolved Oxygen levels around



(a) pH (b) Dissolved Oxygen (c) Specific Conductance (d) Water Temperature

Fig. 3: Spatial interpolation for predicting four water quality indicators in July 2023, including pH (a), Dissolved Oxygen (b), Specific Conductance (c), and Water Temperature (d).



(a) pH (b) Dissolved Oxygen (c) Specific Conductance (d) Water Temperature

Fig. 4: Interpretability of feature variable importances for predicting pH value. The interpretability are exemplified by XGBoost.

10 mg/L, indicating a healthy aquatic environment.

Specific Conductance patterns across California show high values near the San Francisco Bay Area (Fig.3c), possibly due to the colder California Current and natural dilution in the Central Valley and mountains. Near the Bay Area, elevated conductance might result from industrial, urban influences, and seawater erosion.

Water Temperature increases from north to south, aligning with topography (Fig. 3d). Southern California's desert areas reach temperatures around 30°C attributed to lower latitude, desert climate. Conversely, northern California and the Sierra Nevada maintain cooler temperatures around 10°C, due to higher latitudes, altitudes, dense forests. Coastal regions exhibit higher temperatures, likely influenced by human activities. Analyzing water temperature is crucial for understanding climate change, geothermal activities, vegetation shifts.

3.3. Interpretability

In predicting four indicators of surface water quality using the XGBoost model, we demonstrate the proportional impact of nine feature variables on the model's predictions (Fig. 4). For pH and dissolved oxygen, the most significant influences are water temperature and latitude, with each other's influence being the next most significant (Fig. 4a and Fig. 4b). Due to seawater erosion, the electrical conductivity in the bay area is exceptionally high, making longitude and latitude the most critical features affecting the estimation of electrical conductivity. Moreover, seawater erosion is a cumulative process; hence, changes over the years also significantly affect the estimation of electrical conductivity. California's overall water temperature is strongly correlated with the month due to its being a seasonally regular changing water quality indicator. These spatiotemporal variables and water quality indicators have strong, mutually intrinsic relationships. Conversely, the Climate Zone has a minor weight in water quality prediction. Although we incorporated the five major climate classifications (with California having only three major climates) as a method to categorize the climate at the stations, it did not show a dominant role in model predictions, possibly due to the impacts of climate change. The reason for the low significance of Geographical Type is that seawater intrusion's impact on inland waters does not have a clear boundary, which is influenced by factors such as the varying topography and geomorphology of coastlines and seasons.

4. CONCLUSION

In this paper, we extensively evaluated four water quality indicators across the entirety of California over the past 70 years, considering both Climate Zone and Geographical Type. We found that the ML model incorporating spatio-temporal variables and other intrinsically related water quality indicators, Climate Zone, and Geographical Type, can reliably estimate four water quality indicators in California, including pH, Dissolved Oxygen, Specific Conductance, and Water Temperature. We established spatial interpolations for four water quality indicators and visualized their distributions. We observed that the gradient changes of pH, Dissolved Oxygen, and Water Temperature are from north to south. We discovered that Specific Conductance is unusually high in the Bay Area, which might indicate the influence of extensive pollution or seawater erosion. Our future work will dive deeper into other environmental factors that could influence water quality, such as topography, industrial emissions, agricultural activities, and environmental protection policy.

5. REFERENCES

[1] Wisam M Khadra, Ata R Elias, and Michel A Majdalani, "A systematic approach to derive natural background levels in groundwater: Application to an aquifer in north lebanon perturbed by various pollution sources," *Science of The Total Environment*, vol. 847, pp. 157586, 2022.

[2] Arminder Kaur, Siddharth Vats, Sumit Rekhi, Ankit Bhardwaj, Jharna Goel, Ranjeet S Tanwar, and Komal K Gaur, "Physico-chemical analysis of the industrial effluents and their impact on the soil microflora," *Procedia Environmental Sciences*, vol. 2, pp. 595–599, 2010.

[3] Sébastien Sauvé, Sébastien Lamontagne, Jerome Dupras, and Walter Stahel, "Circular economy of water: Tackling quantity, quality and footprint of water," *Environmental Development*, vol. 39, pp. 100651, 2021.

[4] Ken Buesseler, Michio Aoyama, and Masao Fukasawa, "Impacts of the fukushima nuclear power plants on marine radioactivity," *Environmental science & technology*, vol. 45, no. 23, pp. 9931–9935, 2011.

[5] Vincent Rossi, Erik Van Sebille, Alexander Sen Gupta, Véronique Garçon, and Matthew H England, "Multi-decadal projections of surface and interior pathways of the fukushima cesium-137 radioactive plume," *Deep Sea Research Part I: Oceanographic Research Papers*, vol. 80, pp. 37–46, 2013.

[6] Peiyue Li and Jianhua Wu, "Drinking water quality and public health," *Exposure and Health*, vol. 11, no. 2, pp. 73–79, 2019.

[7] Eric Kauffman et al., "Climate and topography," *Atlas of the Biodiversity of California*, vol. 12, pp. 15, 2003.

[8] Michael J. Pennino, Scott G. Leibowitz, Jana E. Compton, Ryan A. Hill, and Robert D. Sabo, "Patterns and predictions of drinking water nitrate violations across the conterminous united states," *Science of the Total Environment*, vol. 722, 2020.

[9] Allison R. Sherris, Michael Baiocchi, Scott Fendorf, Stephen P. Luby, Wei Yang, and Gary M. Shaw, "Nitrate in drinking water during pregnancy and spontaneous preterm birth: A retrospective within-mother analysis in california," *Environmental Health Perspectives*, vol. 129, no. 5, 2021.

[10] Arianna Q Tariqi and Colleen C Naughton, "Water, health, and environmental justice in california: Geospatial analysis of nitrate contamination and thyroid cancer," *Environmental Engineering Science*, vol. 38, no. 5, pp. 377–388, 2021.

[11] Celia Z Rosecrans, Bernard T Nolan, and Jo Ann M Gronberg, "Predicted ph at the domestic and public supply drinking water depths, central valley, california," Tech. Rep., US Geological Survey, 2017.

[12] Paul A Work, Maureen A Downing-Kunz, and Daniel N Livsey, "Record-high specific conductance and water temperature in san francisco bay during water year 2015," Tech. Rep., US Geological Survey, 2017.

[13] Darren L Ficklin, Yuzhou Luo, and Minghua Zhang, "Watershed modelling of hydrology and water quality in the sacramento river watershed, california," *Hydrological processes*, vol. 27, no. 2, pp. 236–250, 2013.

[14] Jingjing Li and Xiaohan Zhang, "Beach pollution effects on health and productivity in california," *International Journal of Environmental Research and Public Health*, vol. 16, no. 11, pp. 1987, 2019.

[15] Houlin Chen and Meredith Franklin, "Spatio-temporal modeling of surface water quality distribution in california (1956-2023)," *arXiv preprint arXiv:2311.12736*, 2023.

[16] California Department of Water Resources, "Water quality data," 2023, Accessed: 2023-07-31.

[17] Maurizio Barbieri, Marino Domenico Barberio, Francesca Banzato, Andrea Billi, Tiziano Boschetti, Stefania Franchini, Francesca Gori, and Marco Petitta, "Climate change and its effect on groundwater quality," *Environmental Geochemistry and Health*, pp. 1–12, 2021.

[18] Hylke E Beck, Niklaus E Zimmermann, Tim R McVicar, Noemi Vergopolan, Alexis Berg, and Eric F Wood, "Present and future köppen-geiger climate classification maps at 1-km resolution," *Scientific data*, vol. 5, no. 1, pp. 1–12, 2018.

[19] Ismail Abd-Elaty and Martina Zelenakova, "Saltwater intrusion management in shallow and deep coastal aquifers for high aridity regions," *Journal of Hydrology: Regional Studies*, vol. 40, pp. 101026, 2022.

[20] Lorenzo Mentaschi, Michalis I Vousdoukas, Jean-Francois Pekel, Evangelos Voukouvalas, and Luc Feyen, "Global long-term observations of coastal erosion and accretion," *Scientific reports*, vol. 8, no. 1, pp. 12876, 2018.

[21] Mohammed S Hussain, Hany F Abd-Elhamid, Akbar A Javadi, and Mohsen M Sherif, "Management of seawater intrusion in coastal aquifers: a review," *Water*, vol. 11, no. 12, pp. 2467, 2019.

[22] Liangqi Yuan, Houlin Chen, Robert Ewing, Erik Blasch, and Jia Li, "Three dimensional indoor positioning based on passive radio frequency signal strength distribution," *IEEE Internet of Things Journal*, vol. 10, no. 15, pp. 13933 – 13944, March 2023.

[23] Liangqi Yuan, Houlin Chen, Robert Ewing, and Jia Li, "Passive radio frequency-based 3d indoor positioning system via ensemble learning," *arXiv preprint arXiv:2304.06513*, April 2023.

[24] Pejman Tahmasebi, Serveh Kamrava, Tao Bai, and Muhammad Sahimi, "Machine learning in geo-and environmental sciences: From small to large scale," *Advances in Water Resources*, vol. 142, pp. 103619, 2020.

[25] M Franklin, O.V. Kalashnikova, and M.J. Garay, "Size-resolved particulate matter concentrations derived from 4.4 km resolution size-fractionated Multi-angle Imaging SpectroRadiometer (MISR) aerosol optical depth over Southern California," *Remote Sensing of Environment*, vol. 196, pp. 312–323, 2017.