Science of the Total Environment 601–602 (2017) 1160–1172

Contents lists available at ScienceDirect

Science of the Total Environment

journal homepage: www.elsevier.com/locate/scitotenv

A hybrid machine learning model to predict and visualize nitrate concentration throughout the Central Valley aquifer, California, USA

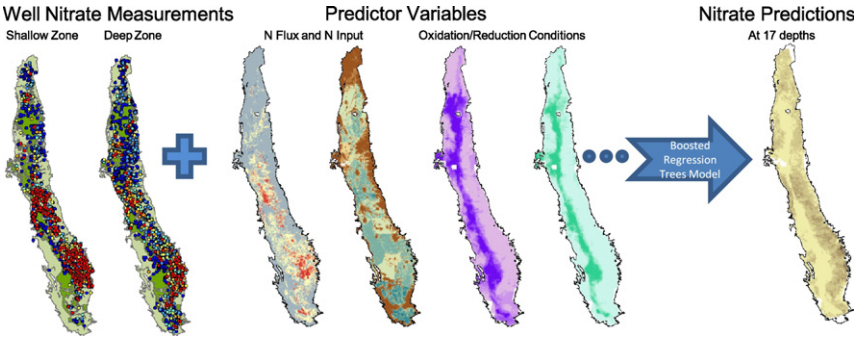
Katherine M. Ransom [a,*], Bernard T. Nolan [b], Jonathan A. Traum [c], Claudia C. Faunt [d], Andrew M. Bell [e], Jo Ann M. Gronberg [f], David C. Wheeler [g], Celia Z. Rosecrans [c], Bryant Jurgens [c], Gregory E. Schwarz [b], Kenneth Belitz [h], Sandra M. Eberts [i], George Kourakos [a], Thomas Harter [a]

[a] University of California, Davis, Department of Land, Air, and Water Resources, United States
[b] U.S. Geological Survey National Water Quality Program, Reston, VA, United States
[c] U.S. Geological Survey California Water Science Center, Sacramento, CA, United States
[d] U.S. Geological Survey California Water Science Center, San Diego, CA, United States
[e] University of California, Davis, Center for Watershed Sciences, United States
[f] U.S. Geological Survey California Water Science Center, Menlo Park, CA, United States
[g] Virginia Commonwealth University, Department of Biostatistics, Richmond, VA, United States
[h] U.S. Geological Survey New England Water Science Center, Northborough, MA, United States
[i] U.S. Geological Survey Ohio Water Science Center, Columbus, OH, United States

HIGHLIGHTS

- Boosted regression tree model produced 3D map of nitrate concentration.
- Hybrid multi-modeling approach used numerical model outputs as predictors.
- Redox characteristics and field scale unsaturated zone N flux were most important.
- Nitrate concentrations <2 mg/L NO_3-N generally conformed to basin subregion.
- Nitrate concentrations >10 mg/L NO_3-N most common in eastern alluvial fans subregion

GRAPHICAL ABSTRACT



ARTICLE INFO

Article history:
Received 6 March 2017
Received in revised form 19 May 2017
Accepted 20 May 2017
Available online 9 June 2017

Editor: D. Barcelo

Keywords:
Groundwater
Nitrate
Boosted regression trees

ABSTRACT

Intense demand for water in the Central Valley of California and related increases in groundwater nitrate concentration threaten the sustainability of the groundwater resource. To assess contamination risk in the region, we developed a hybrid, non-linear, machine learning model within a statistical learning framework to predict nitrate contamination of groundwater to depths of approximately 500 m below ground surface. A database of 145 predictor variables representing well characteristics, historical and current field and landscape-scale nitrogen mass balances, historical and current land use, oxidation/reduction conditions, groundwater flow, climate, soil characteristics, depth to groundwater, and groundwater age were assigned to over 6000 private supply and public supply wells measured previously for nitrate and located throughout the study area. The boosted regression tree (BRT) method was used to screen and rank variables to predict nitrate concentration at the depths of domestic and public well supplies. The novel approach included as predictor variables outputs from existing physically based models of the Central Valley. The top five most important predictor variables included two oxidation/

* Corresponding author.
E-mail address: kmlockhart@ucdavis.edu (K.M. Ransom).

http://dx.doi.org/10.1016/j.scitotenv.2017.05.192