Saint Mary's University GeoSpatial Services

WMC Webinar

Springs and Seeps

EVALUATION OF PREVALENCE PROBABILITY PREDICTION OF GROUND WATER DISCHARGE SPRINGS IN MO USING THE MAXENT MACHINE LEARNING ALGORITHM

Objectives

1. To asses the ability of geospatial and machine learning based modeling to predict presence of ground water discharge springs
2. Produce a predictive model that can be applied to a broader area of interest in locating areas where ground water springs are more likely to occur

Introduction

Remote sensing, geospatial modeling and machine learning have been successfully used to spatially model where ground water discharge occurs (Howard & Merrifield, 2010; Pourtaghi & Pourghasemi, 2014; Gerlach et al., 2022).

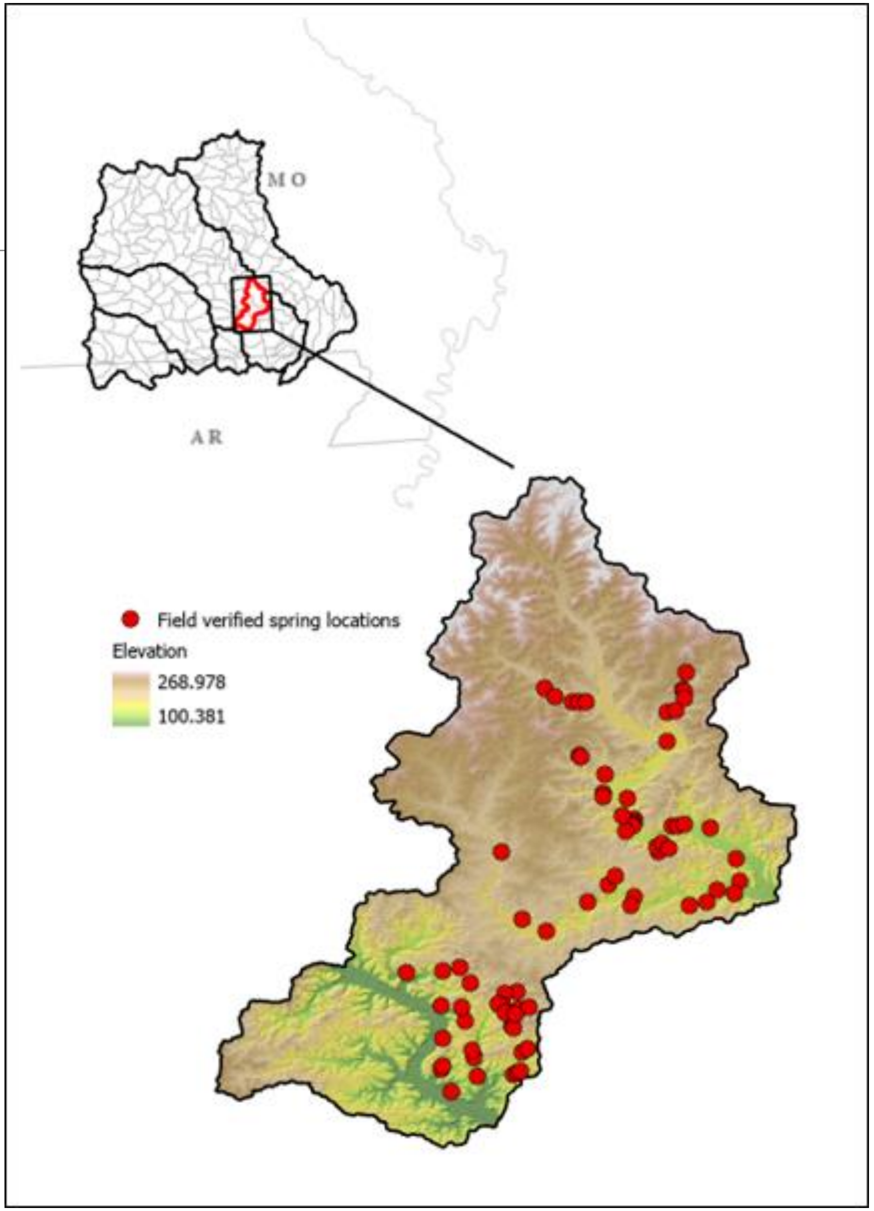
Gerlach et al. (2022) determined that locations of groundwater discharge in the form of springs and seeps could be accurately predicted using a MaxEnt machine learning model and combinations of topographic predictor variables derived from high-resolution lidar digital elevation models (DEM).

We evaluated performance of the six predictor variables identified by Gerlach et al. (2022) in locating springs using a MaxEnt machine learning model. We also identified 10 additional variables that were commonly used in geospatial modeling of ground water discharge and evaluated them as potential predictor variables in the analysis.

Study Area

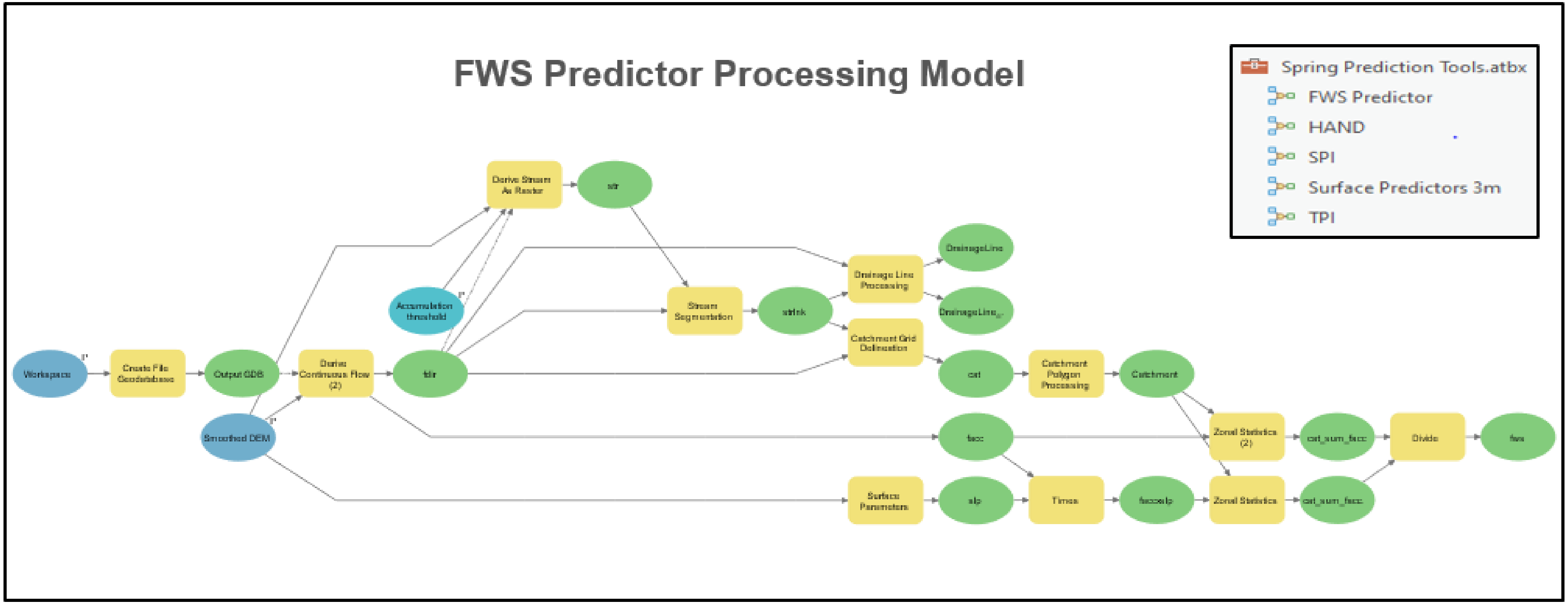
Selection of HUC 12 watersheds for the project was limited to areas that contained an adequate number of field-verified springs from the MO springs database.

The examined watersheds collectively contained a total of 76 field-verified springs which provided the largest number of clustered observations for modeling.



Saint Mary's University
GEOSPATIAL SERVICES

Data Assembly



Modeling: MaxEnt

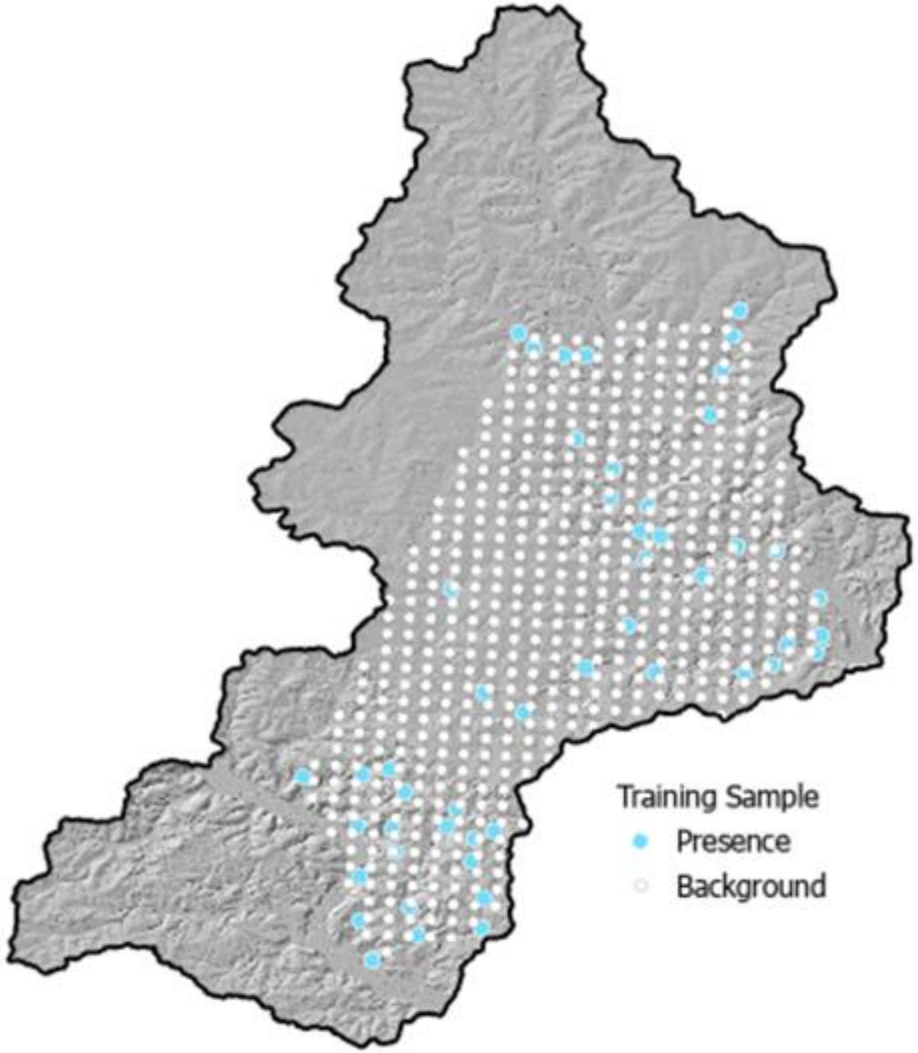
MaxEnt modeling has been used extensively in species distribution modeling and is ideal for this analysis because of its ability to work with presence-only observations and small sample sizes (Phillips et al., 2006).

Also, MaxEnt is a machine learning algorithm that has the capability to use basis functions to transform explanatory variables, allowing more complex modeling of relationships between explanatory variables and the dependent variable being predicted. A process called regularization is used to identify the input and transformed variables that are most significant for prediction. This process also removes data redundancy in the model, thereby addressing the issue of potential multicollinearity of explanatory variables (Feng et al., 2019).

Model Training

Model training parameters for MaxEnt modeling included: basis functions (Linear, Quadratic, Product and Hinge) for variable transformation, spatial thinning with a minimum distance of 500m, a value of 100 for relative weight of presence to background, C-log-log probability transformation, and K-Fold validation using three groups.

For each training model, presence probability cutoff values of 0.4 and 0.5 were specified for model validation.



Saint Mary's University
GEOSPATIAL SERVICES

Model Training

The omission rate identifies the percent of presence observations misclassified.

AUC quantifies the ability of the model to distinguish between presence and absence of springs. A value of 0.5 for AUC represents a model that is no better than random and a value of 1 is a perfect scoring model.

K-fold validation divides the sample data into three groups. The model iterates through each group with one group reserved for validation and the other two groups used for training the model. The K-fold average is the average percent correct for these three runs of the model. K-fold average represents the ability of the model to predict presence of springs in unknown locations.

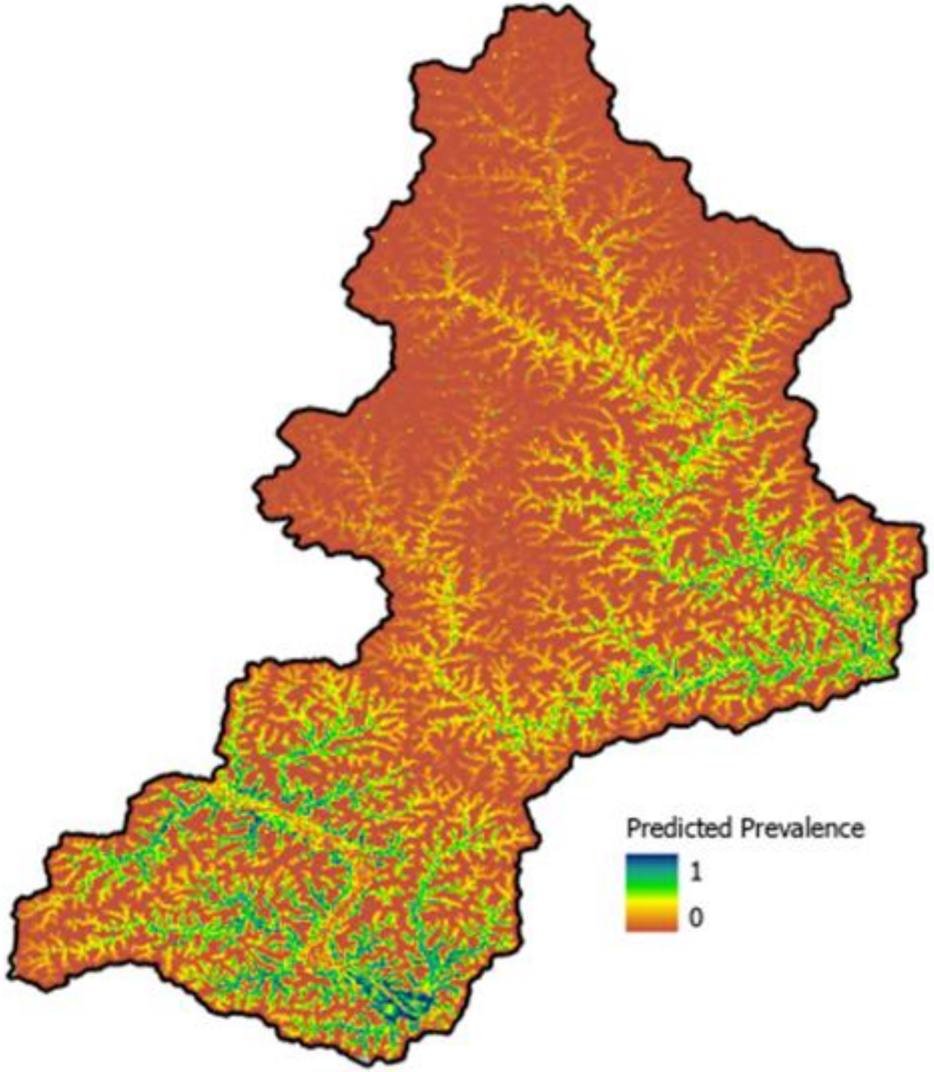
Model	Omission Rate	AUC	K-fold Avg	Prevalence cutoff	Functions	Number of points	Minimum distance
16 var	0.16	0.96	58.9	0.5	all	49	500
16 var	0.10	0.96	56.9	0.4	all	49	500
10 var	**0.18**	**0.95**	**67.2**	**0.5**	**all**	**49**	**500**
10 var	0.10	0.94	76.0	0.4	all	49	500
6 var	0.22	0.91	66.8	0.5	all	49	500
6 var	0.18	0.92	77.1	0.4	all	49	500

Saint Mary's University
GEOSPATIAL SERVICES

Final Model

The 10-variable model produced the best validation results at the 0.5 prevalence cutoff value. This final model was used to predict the prevalence probability of springs for the entire pilot study area.

The prevalence probability raster surface contains values from 0 to 1, with 0 being low probability that the cell will contain a spring and 1 being a high probability.



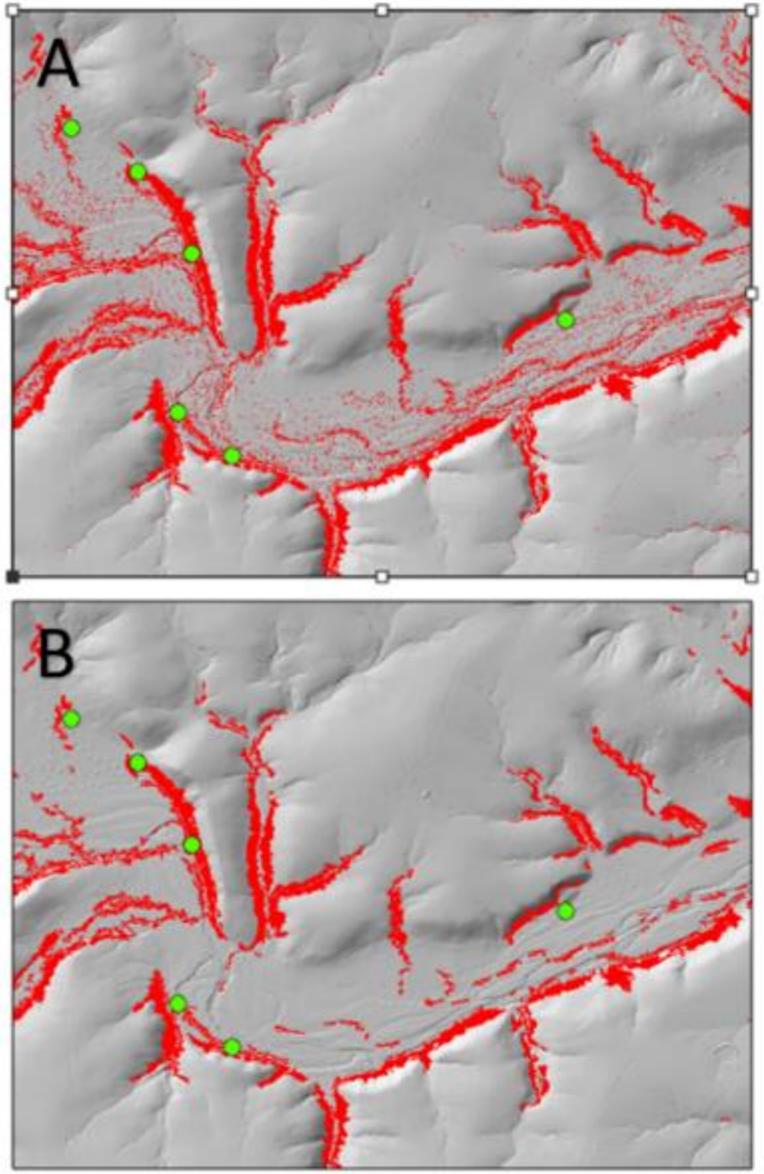
Saint Mary's University GEOSPATIAL SERVICES

Accuracy Assessment

For accuracy assessment of the final model, an independent testing dataset was assembled which consisted of a total of 81 points from the karst fen and MO springs databases.

The prevalence probability surface was then classified using 0.75 and 0.5 cutoff values for the accuracy assessment.

These results were also filtered to remove smaller areas less than 100 square meters to remove noise and improve the interpretability of the output.



Saint Mary's University
GEOSPATIAL SERVICES

Accuracy Assessment

Accuracy was assessed by examining proximity of the classified prevalence probabilities to the independent testing sample points. Proximity was evaluated for testing samples by counting the number of sample points within 10 or 20 meters of the classified probability results.

The filtered and unfiltered classified probabilities all performed well (percent correct greater than 70%) when evaluating proximity within a 20m distance of the reference samples.

Results indicate the filtered classified probabilities are less accurate but the tradeoff is increased interpretability of the final classified results.

	Prevalence threshold	10m count	10m %	20m count	20m %	Number of Samples
10 var filtered	0.75	46	0.57	59	0.73	81
10 var unfiltered	0.75	66	0.81	76	0.94	81
10var filtered	0.5	68	0.84	74	0.91	81
10 var unfiltered	0.5	73	0.90	79	0.98	81

Saint Mary's University
GEOSPATIAL SERVICES

Summary

1. Training validation and the independent accuracy assessment indicates that MaxEnt modeling has the potential to be a useful tool for locating ground water discharge springs.
2. The prevalence probabilities produced by the MaxEnt algorithm do not provide specific locations of where springs occur but can provide a sampling frame that identifies areas where springs are more likely to occur, thus potentially minimizing the amount of field investigation required to locate springs.
3. Detailed examination of the results indicate that the model assigns higher prevalence probabilities to areas with high slope bordering streams/rivers, at mid to lower elevation, where there is higher flow accumulation and stream/river convergence. Typically, these areas also have abrupt changes in topography represented by topographic roughness and curvature values. These results are consistent with other findings in the reviewed literature for modeling of ground water discharge using topographic variables.

Recommendations

Results could be further improved with a more rigorous assessment and selection of significant explanatory variables. Incorporation of variables related to climate and geology may also improve model results if available.

Also, additional machine learning algorithms such as Random Forest could be evaluated to determine if a more accurate model can be produced.

Finally, trained models should be tested to assess model ability to predict prevalence of springs in different geographic areas outside of the pilot study area.

Saint Mary's University
GEOSPATIAL SERVICES

References

Howard, J.; Merrifield, M. Mapping Groundwater Dependent Ecosystems in California. PLoS ONE 2010, 5, e11249

Pourtaghi, Z.S.; Pourghasemi, H.R. GIS-based groundwater spring potential assessment and mapping in the Birjand Township, southern Khorasan Province, Iran. Hydrogeol. J. 2014, 22, 643–662.

Gerlach, M.E.; Rains, K.C.; Guerrón-Orejuela, E.J.; Kleindl, W.J.; Downs, J.; Landry, S.M.; Rains, M.C. Using Remote Sensing and Machine Learning to Locate Groundwater Discharge to Salmon-Bearing Streams. *Remote Sens.* 2022, *14,* 63.

Arabameri, A., Pal, S. C., Rezaie, F., Nalivan, O. A., Chowdhuri, I., Saha, A., Lee, S., & Moayedi, H. (2021). Modeling groundwater potential using novel GIS-based machine-learning ensemble techniques. Journal of Hydrology: Regional Studies, 36, 100848.

Phillips, Steven J., Robert P. Anderson, Robert E. Schapire. 2006. "Maximum entropy modeling of species geographic distributions." Ecological Modelling, 190: 231-259.

Feng, X., Park, D.S., Liang, Y., Pandey, R., Papeş, M. 2019. Collinearity in ecological niche modeling: Confusions and challenges. Ecol Evol. Aug 20;9(18):10365-10376.