Random forest for identification and characterization of groundwater dependent ecosystems

I. C. Perez Hoyos, N. Krakauer & R. Khanbilvardi
NOAA Cooperative Remote Sensing Science and Technology (NOAA CREST Center), City College of New York, USA

Abstract

Anthropogenic actions such as groundwater pumping, agricultural practices, industrialization, and waste disposal can greatly affect groundwater resources which would eventually drive changes in vulnerable ecosystems. Therefore, it is clear that there is a need to identify the locations of groundwater dependent ecosystems (GDEs) to enable the development of policies that adequately address their protection. The purpose of this study is to propose a method based on geospatial data sets and random forest algorithm to map the distribution of GDEs in the United States at 1 km spatial resolution. This paper presents the results in Nevada. The method is based on the principle that ecosystems will use water in proportion to its availability and the dependence on that resource will be expected to increase with higher aridity of the environment. Results show that random forest is a promising technique for the identification and characterization of GDEs using geospatial data sets as predictor variables.
Keywords: groundwater dependent ecosystems, random forest, overlay analysis, water table depth, aridity.

1 Introduction

Groundwater Dependent Ecosystems (GDEs) are plants, animals, and other organisms that depend on groundwater to maintain their structure and composition, as well as to sustain their life processes. There are several types of GDEs, but they all depend on the surface or subsurface expression of groundwater. The main categories of GDEs include the following [1–3]. Terrestrial vegetation (phreatophytes) and fauna, baseflow in river systems, ecosystems in streams and

WIT Transactions on Ecology and The Environment, Vol 196, © 2015 WIT Press
www.witpress.com, ISSN 1743-3541 (on-line)
doi:10.2495/WRM150081

lakes fed by groundwater, springs and seeps, wetlands, aquifers, karst, and cave ecosystems. The degree of dependence on groundwater is variable and ecosystems might be completely reliant or they might require groundwater only a few months of the year [3]. GDEs are of crucial importance for a variety of ecological resources such as terrestrial vegetation, wetlands, wildlife, sensitive fish, and other organisms that are highly vulnerable to be affected by variations in groundwater [1].

In the current conditions, it is clear that there is a need to place restrictions in the amount of groundwater that can be extracted from an aquifer. The first step in the conservation of GDEs is to map their distribution, and then characterize the degree of dependence to be able to predict the ecological response to changes in water supply. Information about the location of GDEs is not readily available in the United States and only a few efforts to map the location and extent of GDEs at a national scale have been undertaken in countries like Australia and South Africa. The purpose of this study is to propose a method based on geospatial data sets and Random Forest (RF) algorithm to map the distribution of GDEs in the United States at 1 km spatial resolution. This paper presents the results of the application of this method in Nevada. The method is based on the principle that ecosystems will use water if the resource is available, and if that resource is limited, the ecosystem will create a functional dependence based on the spatiotemporal availability of the resource [4]. That dependence is expected to increase with greater aridity of the associated environment [5]. Water table position serves the purpose of ecological filter by regulating moisture [6]. For this reason, this parameter will be used as a proxy for the defining the location of GDEs.

2 Background

Previous efforts worldwide have focused in the development of feasible, consistent, and cost-effective techniques to map GDEs at a local, regional, and even national scale. In Australia, it was not until the early 1970s that the approach to manage water resources by solely considering human needs was questioned. The consequences of groundwater overdraft on ecosystems were documented and the need to mitigate environmental impacts arising from these practices became obvious. In 1994 the Australian Government established a set of reforms to strive for a sustainable water industry [7]. Several efforts to map the location of GDEs in Australia have been undertaken. The most relevant to this study is the development of a GDE Atlas as part of a National Water Commission project which consists in a spatially based tool depicting the location of potential GDEs in the country [8]. Governmental agencies in the United States have been investing more resources into the development of strategies to identify the location. An inventory field guide was created with the purpose of providing a national protocol for collecting ground-based data that can serve as the basis for defining the location, extent, and characteristics of springs and wetlands on a local scale.

3 Methods

3.1 Study area

The study site for this project is Nevada, a state located in the south western region of the United States. Nevada is mainly formed by desert and semiarid climate regions and it is characterized by sudden changes in elevation, with the presence of both narrow mountains and flat arid valleys. Nevada was selected as the study area because of its unique geomorphological features and wide range of climatic conditions that have led to the presence of a great variety of vegetation types.

3.2 Water table depth

From the many factors influencing GDEs, depth to water table is probably the most important one. In ecosystems that are dependent on the subsurface expression of groundwater, the depth that roots must reach to access groundwater is a major limitation concerning their capacity to use the resource [9]. Schenk and Jackson [10] found that the vast majority of plant species develop about 95% of their root biomass in the shallower 2.0 m of soil. Water table depth was selected as the single most important indication of groundwater availability, and therefore areas where water table depth is shallow have a greater potential to develop an interaction with terrestrial ecosystems. In this study, depth to groundwater (measured in feet) was obtained from the National Water Information System (NWIS) [11], which contains records from 1960 to 2000 in 8210 different points in Nevada. These observations may represent one observation at a time or a mean value obtained from a time series of groundwater level measurements (see Figure 1).

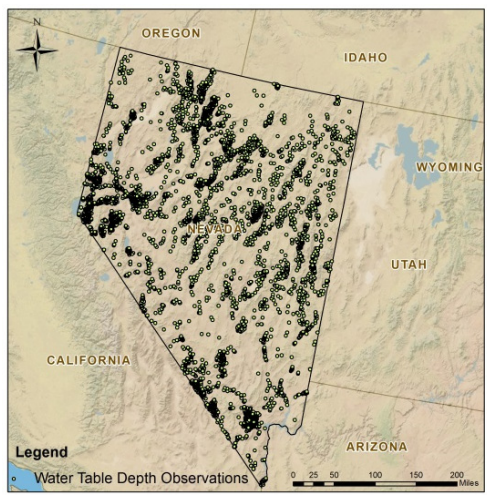

Figure 1: Depth to groundwater records obtained from the National Water Information System (NWIS) for the state of Nevada, USA.

3.3 Random forest regression

The method that will be used to predict water table depth using climate, topography, and vegetation as predictors is random forest. RF is an ensemble learning algorithm developed by Breiman [12] that fits many classification or regression trees (CART) models to random subsets of the input data and uses the combined decision trees (forest) for prediction. Important features of the RF are:

- RF estimates the importance of the predictor variables when modelling the response variable using permutation accuracy.
- RF has the ability to identify the proximity between pairs of data points which can be useful for understanding the structure of the data, clustering, and locating outliers [13].
- RF does not suffer from overfitting because the amount of trees that are grown is large, resulting in a limited generalization error (true population error) [14].

The algorithm for RF consists in building a forest of uncorrelated trees. This is accomplished by using bootstrap aggregation in which given a training data set, random samples with replacement are selected and trained using decision or regression trees. Each individual tree is grown using a randomized subset of predictor variables. The trees are grown to the largest extent possible without pruning, and they are aggregated by averaging them. Out-of-bag (OOB) samples are used to calculate variable importance and to get an unbiased estimate of the test set error which is one of the advantages of RF because there is no need for cross-validation. In this study, the implementation of the RF algorithm within a GIS interface was performed using the Marine Geospatial Ecology Tools (MGET) [15]. MGET implements the classic RF algorithm using the RF package available in the environment for statistical computing known as R [12]. The number of trees that are used is 500. More trees have the ability yield models with greater stability and covariate importance estimates. However, computer resources required also increase. It is suggested that 500 trees or more are used for large datasets [15]. In this case, the number of predictor variables is low, therefore 500 trees are considered sufficient. On the other hand, the number of variables tried at each split is 2, since this value is suggested to be set equal to a third of the number of predictor variables [16].

3.3.1 Predictors

These are the data sets to which the model is fitted. In this study, all predictor variables selected are treated as continuous variables.

3.3.1.1 Topography It is clear that GDEs are typically found in locations where groundwater is known or expected to be shallow (e.g. topographically low areas and major breaks of topographic slope) [17]. Several authors have documented the ecosystem response to landforms and how the composition and structure of vegetation is determined by geomorphic events [18, 19]. Landforms, usually characterized by topography and geology, deeply influence spatial variations in ecological variables such as water availability and exposure to radiant solar energy. Topography is linked with climate through varying heights and degree of

ground-surface inclination, controlling the intensity of key factors (such as hydrology) that are important to plants and to the soil-forming processes [20]. Land surface form is assessed with variables such as elevation, slope, and aspect. Slope is a measure of the steepness of the surface at a particular location. On the other hand, the aspect is a measure of the direction of steepest slope for a location on the surface. It is typically measured in degrees and it is estimated by assigning a number, between 0 and 360, to each cell in the grid depending on the direction that the cell faces. Characteristics such as ground moisture, snow retention, vegetation, and surface temperature are all deeply influenced by aspect. Elevation was obtained from the National Elevation Dataset (NED) developed by the US Geological Survey at 30 m spatial resolution. This Digital Elevation Model (DEM) was also used as the basis for the calculation of slope and aspect.

3.3.1.2 Climate Water table depth fluctuates in response to precipitation events because groundwater is recharged by precipitation that percolates through soils. Fan *et al.* [6] found that at the regional scale, climate is the dominant driver for water table position simulation, whereas at more local scales the primary driver is topography. They also emphasized the fact that when solar radiation is not limiting, the distribution of vegetation is associated with moisture gradients given by rainfall patterns. On the other hand, groundwater dependency by terrestrial vegetation is directly linked with the water budget. If the total amount of water that is being used by terrestrial vegetation in a given site for a specific time period can be demonstrated to be considerably larger than the total precipitation for the site, and there is no significant lateral flow, it can be concluded that this ecosystem depends to a certain degree on groundwater [3]. For the above reasons, climatic variables such as mean daily Precipitation in cm (PRECIP), mean daily maximum Temperature in degrees Celsius (TEMP), and mean daily Shortwave Radiation in $MJ/m^2/day$ (S_RAD) were considered predictors of the water table depth. These parameters were obtained from Daymet. These are model-produced estimates, at 1 km resolution, of daily weather parameters (temperature, precipitation, humidity, and radiation) based on daily meteorological observations. The required model inputs include a DEM and ground-based observations of temperature and precipitation [21]. The datasets used in this study represent daily climatological summaries produced on a 1 km grid for the period of record of 1980 to 1997.

3.4 Definition of aridity

The Aridity Index (AI) is a numerical value that indicates the degree of dryness of the climate in a given place. This indicator is used to estimate availability of precipitation over atmospheric water demand. The United Nations Environment Programme (UNEP) adopted an index defined as Mean Annual Precipitation (MAP) divided over the Mean Annual Potential Evapotranspiration (PET). In this study, AI is obtained from the Global Aridity Index geodatabase [22] (see Figure 2). This data set is provided at 30 arc seconds (~ 1 km at the equator) and includes year 1950 to 2000. It was produced using MAP values obtained from the WorldClim Global Climate Data and monthly PET (modelled using Hargreaves method) aggregated to annual average values. The AI categories are defined as

follows: AI <0.03 – Hyper arid; AI between 0.03 and 0.2 – Arid; AI between 0.2 and 0.5 – Semi-arid; AI between 0.5 and 0.65 – Dry sub-humid; AI>0.65 – Humid.

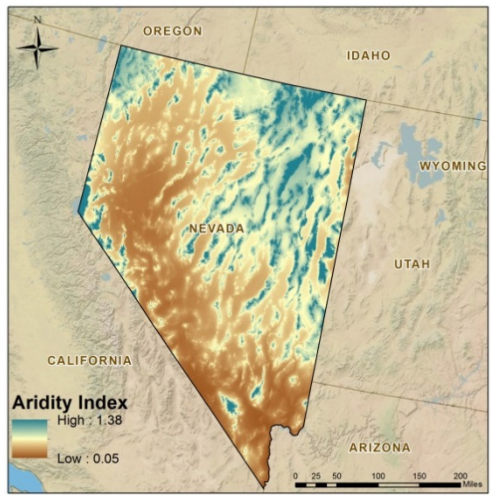

Figure 2: AI map for Nevada, USA. Low values indicate high aridity whereas high values indicate humid climate.

3.5 Weighted overlay analysis

In order to determine the potential of an ecosystem to be groundwater dependent, the depth to water table and aridity index maps are combined into an integrated model using an overlay analysis. For this study, the weighted overlay analysis technique available in ArcGIS Spatial Analyst extension is used. The technique has the advantage of providing means to prioritize the factors in the analysis. In this study, input layer importance is defined as being equal for the water table depth and the AI, since they are both considered equally relevant for determining groundwater dependence. The weighted overlay function combines the two raster layers multiplying each by their weight and adding them together.

4 Results

4.1 Variable importance

The machine learning algorithm used to predict water table depth is RF which estimates the variable importance using two methods, based in Mean Squared Error (MSE) or the total reduction in sum squares. For this analysis, the importance of each predictor variable was estimated using MSE and it is performed by calculating the difference between the MSE of the whole model and the MSE that represents the prediction accuracy of the OOB part of the data after permuting each predictor variable [14]. The resulting plot is shown in Figure 3. IncMSE represents the percentage increase in accuracy (calculated using mean square

errors). The greater the value the more important the predictor variable is. Mean daily maximum air temperature is the most important predictor of water table depth. This can be considered a relevant finding because air temperature has a direct effect on the thermal regime of a given location, and therefore in the actual evapotranspiration and distribution of the precipitation, which in turn influence water table position. If this variable was omitted, the quality of the model could be reduced. Elevation was also found to be a relevant predictor of water table depth, followed by daily shortwave radiation and precipitation. Slope and aspect have the smallest influence on the quality of the model. Therefore, it can be concluded that the distribution of water table position is more strongly driven by climate as compared to topography.

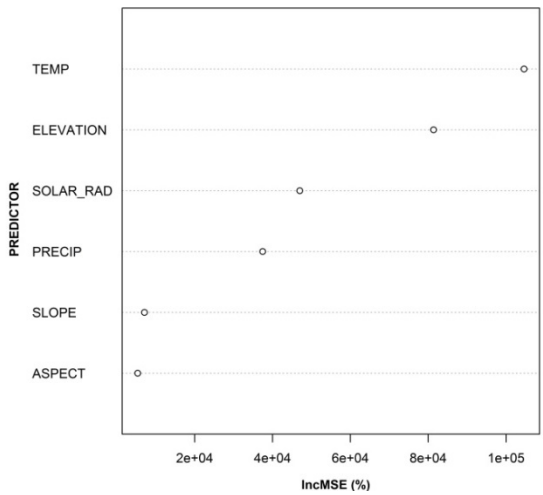

Figure 3: Variable importance plots for predictor variables from RF regression.

4.2 Partial dependence

Partial dependence plots are also used to comprehend the relationships between the individual predictor variables and the response variable (water table depth observations) obtained by implementing RF. In Figure 4 the influence of each predictor on water table (vertical axis) is estimated when the remaining predictors are kept constant. The partial dependence plot for TEMP resembles what was expected. Because it is an important predictor of the response variable, the predictor variable TEMP shows partial dependence over its entire range (0 to 30 degrees Celsius) and a wide range of water table depth values (0 to 450 ft). The partial dependence of ELEVATION varies randomly around 200 ft over its entire range. On the other hand, predictor variable SOLAR_RAD displays partial dependence only in a portion of its range (between 15 and 20 $MJ/m^2/day$). As it was expected, the partial dependence plots for the least important variables are almost horizontal lines around the mean of water table depth (see Figure 4).

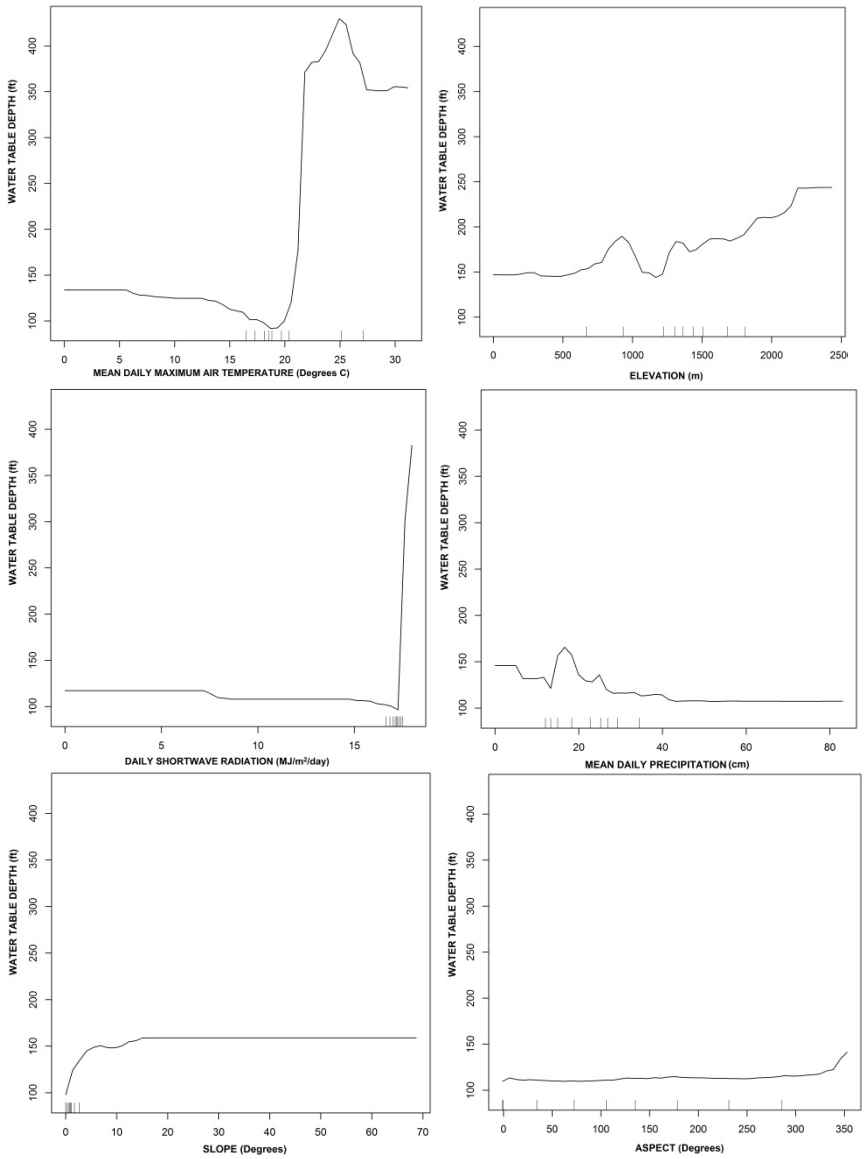

Figure 4: Partial dependence plots for the six selected predictor variables of water table depth using RF regression.

4.3 Predicted water table depth

After implementing the RF algorithm, the output model is used to compute water table depth continuously in the state of Nevada. The input data are the raster layers depicting the predictor variables. Before implementing the model, the topography

input layers (ELEVATION, SLOPE, ASPECT) are upscaled from 30 m to 1 km spatial resolution, which is the desired resolution for the predicted water table depth as well as the resolution for the climate layers. After the analysis is performed, the resulting map shows water table position (in feet) at all points at 1 km spatial resolution (see Figure 5).

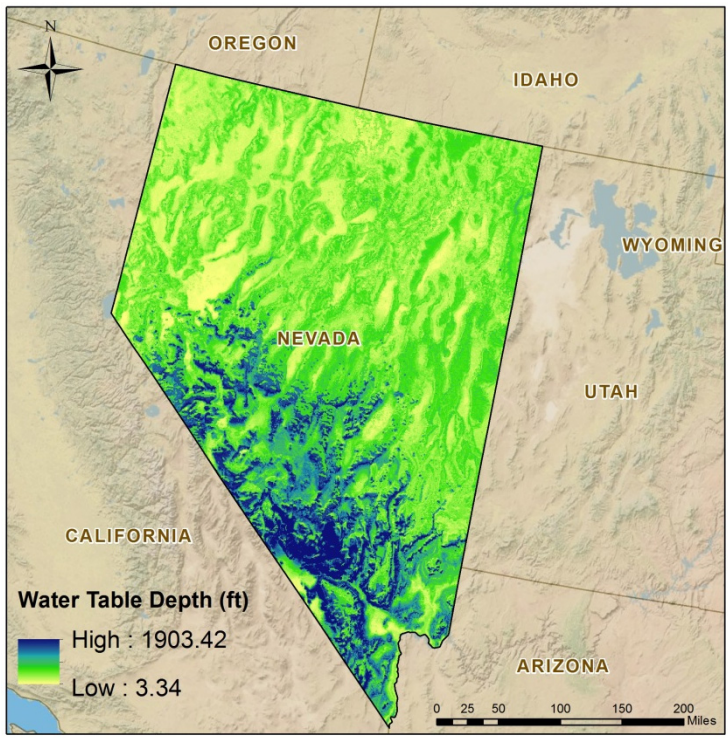

Figure 5: Water table depth (ft) for the state of Nevada as predicted by RF algorithm.

4.4 Groundwater dependence potential map

Weighted overlay analysis is used for the purpose of producing a map that depicts the potential of an ecosystem to be groundwater dependent. This analysis is based on the fact that ecosystems will use resources in proportion to their availability, and if the resource (in this case groundwater) is available, the degree of this reliance will depend on the aridity. Hence, a low value of AI (hyper arid climate) and a low value of predicted WTD (shallow water table) will result in a low value in the output of the overlay analysis, which will indicate high potential of an ecosystem to be groundwater dependent (see Figure 6). A more humid climate, in which the ecosystems' water requirements are met by precipitation in addition to a deep water table, will indicate that there is a low likelihood of the ecosystem to be groundwater dependent (as depicted in green in Figure 6).

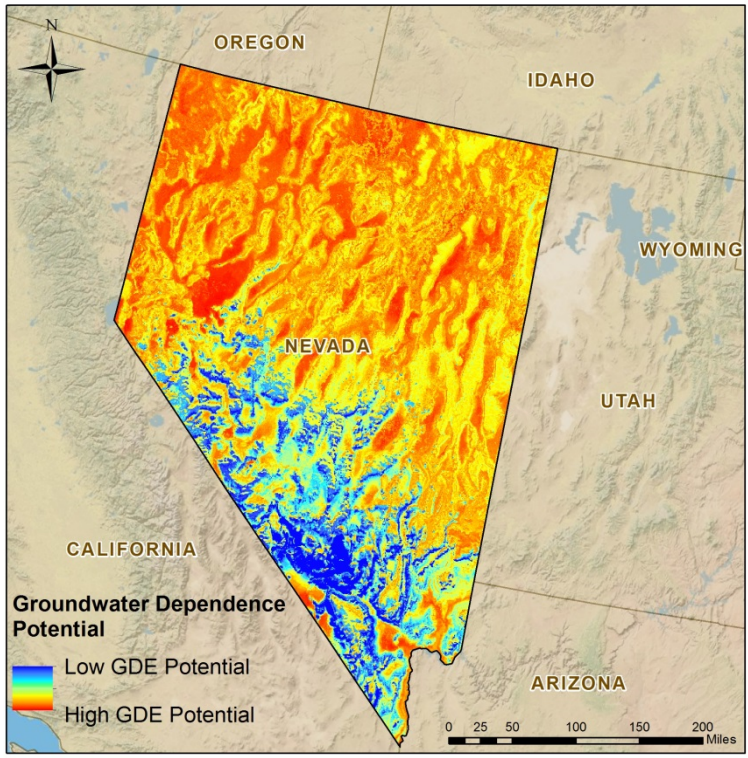

Figure 6: Map of groundwater dependence potential estimated using integrated water table depth and aridity index using overlay analysis.

5 Conclusions

This large scale analysis represents the first attempt to determine the distribution of GDEs at such resolution. The model proposed here can also be applied to other areas where information of GDEs or the link between groundwater and ecosystems wants to be understood and characterized. This type of information is crucial in the conservation of biodiversity that could be potentially affected by global-change type droughts. It is also significant in the development of environmental policies or regulations that address ecosystems and groundwater sustainability issues as well as an information tool to depict the extensive distribution of GDEs throughout the country. The use of geospatial datasets is relevant in an endeavour like identifying the location of GDEs at a large scale because ground-based methods can be expensive, time consuming, and labour intensive. RF is a promising technique in the process of mapping GDEs because of their ability to provide meaningful analysis of nonlinear and complex variables such as the ones found in hydro-ecological studies. Future work will include the use of remote sensing datasets such as Normalized Difference Vegetation Index (NDVI) and

Land Surface Temperature (LST) as additional predictor variables, as well as the validation of the resulting groundwater dependence potential map using datasets that depict groundwater discharge areas and regions where GDEs have been previously identified.

Acknowledgements

The authors gratefully acknowledge support from NOAA under grants NA11SEC4810004 and NA12OAR4310084. All statements made are the views of the authors and not the opinions of the funding agency or the US government.

References

[1] Glasser, S., Gauthier-Warinner, J., Keely, J., Tucci, P., Summers, P., Wireman, M., & McCormack, K., Technical Guide to Managing Ground Water Resources, 2007.
[2] Howard, J. & Merrifield, M., Mapping groundwater dependent ecosystems in California, *PLoS One*, vol. 5, no. 6, 2010.
[3] Eamus, D., Identifying groundwater dependent ecosystems: A guide for land and water managers, Sydney, 2009.
[4] Hatton, T. & Evans, R., Dependence of Ecosystems on Groundwater and its Significance to Australia, *LWRRDC*, Occasional Paper no. 12, 1998.
[5] Dresel, P. E., Clark, R., Cheng, X., Reid, M., Terry, A., Fawcett, J. & Cochrane, D., Mapping Terrestrial Groundwater Dependent Ecosystems : Method Development and Example Output, Melbourne, 2010.
[6] Fan, Y., Li, H. & Miguez-Macho, G., Global patterns of groundwater table depth., *Science*, vol. 339, no. 6122, pp. 940–943, 2013.
[7] Eamus, D. & Froend, R., Groundwater-dependent ecosystems: the where, what and why of GDEs, *Aust. J. Bot.*, vol. 54, pp. 91–96, 2006.
[8] Sinclair Knight Merz, Australian groundwater dependent ecosystems toolbox part 1: assessment framework, Canberra, Australia, 2011.
[9] Gow, L., Brodie, R. S., Green, R., Punthakey, J., Woolley, D., Redpath, P. & Bradburn, A., Identification and monitoring GDEs using MODIS time series: Hat Head National Park – a case study, *Groundwater 2010: The challenge of sustainable management*, 2010.
[10] Schenk, H. J. & Jackson, R. B., The global biogeography of roots, *Ecol. Monogr.*, vol. 72, no. 3, pp. 311–328, 2002.
[11] United States Geological Survey, USGS Water Data for the Nation, 2015. Online. http://waterdata.usgs.gov/nwis/.
[12] Breiman, L., Random forests, *Mach. Learn.*, pp. 5–32, 2001.
[13] Liaw, A., Wiener, M., Classification and Regression by random Forest, *UNC Publ.*, vol. 2, no. December, pp. 18–22, 2002.
[14] Prasad, A. M., Iverson, L. R. & Liaw, A., Newer classification and regression tree techniques: Bagging and random forests for ecological prediction, *Ecosystems*, vol. 9, no. 2, pp. 181–199, 2006.

[15] Roberts, J. J., Best, B. D., Dunn, D. C., Treml, E. A. & Halpin, P. N., Marine Geospatial Ecology Tools: An integrated framework for ecological geoprocessing with ArcGIS, Python, R, MATLAB, and C++, *Environ. Model. Softw.*, vol. 25, no. 10, pp. 1197–1207, 2010.
[16] Breiman, L. & Cutler A., Manual for Setting up, Using, and Understanding Random Forest, 2003.
[17] Sinclair Knight Merz, Environmental Water Requirements to Maintain Groundwater Dependent Ecosystems, Commonwealth of Australia, Canberra, 2001.
[18] Swanson, F. J., Kratz, T. K., Caine, N. & Woodmansee, R. G., Landform Effects on Ecosystem Paterns and Processes, *Bioscience*, vol. 38, no. 2, pp. 92–98, 1998.
[19] Zhou, Y., Wenninger, J., Yang, Z., Yin, L., Huang, J., Hou, L., Wang, X., Zhang, D. & Uhlenbrook, S.,Groundwater–surface water interactions, vegetation dependencies and implications for water resources management in the semi-arid Hailiutu River catchment, China – a synthesis, *Hydrol. Earth Syst. Sci.*, vol. 17, no. 7, pp. 2435–2447, 2013.
[20] Bailey, R. G., Role of Landform in Differentiation of Ecosystems at the Mesoscale (Landscape Mosaics), 2004.
[21] Thornton, P.E., Thornton, M.M., Mayer, B.W., Wilhelmi, N., Wei, Y. & Devarakonda, R., Daymet: Daily Surface Weather Data on a 1-km Grid for North America, Version 2, *Oak Ridge National Laboratory Distributed Active Archive Center*, 2014. Online. http://daac.ornl.gov/.
[22] Zomer, R. J., Trabucco, A., Bossio, D. A., & Verchot, L. V., Climate change mitigation: A spatial analysis of global land suitability for clean development mechanism afforestation and reforestation, *Agric. Ecosyst. Environ.,* vol. 126, no. 1–2, pp. 67–80, 2008.