ECOAQUA
MODELING AND MONITORING OF AN EXPLOITED AQUIFER SYSTEM IN NORTHERN BAJA CALIFORNIA, MEXICO

Christine Schottmüller[1], Björn Riedel[1], Anika Riedel[1], Markus Gerke[1], Wolfgang Niemeier[1]

[1] Technische Universität Braunschweig, Institute for Geodesy and Photogrammetry, Bienroder Weg 81, 38106 Braunschweig (c.schottmueller @tu-bs.de)

Key words: *Socio-hydrologic model*; *Support Vector Machines*; *InSAR*

ABSTRACT

The progressive urbanization, industrialization and cultivation of the Ensenada city region located at the west coast of Mexico causes high pressure on the local and coastal ecosystems. High demand on the freshwater resources in combination with the semi-arid climatic conditions has led to overuse of local aquifers in recent decades. Resulting scarcity and salinization problems, landform changes accompanying infrastructure damage, degradation of soil quality and hydrological droughts present significant challenges to both civil society and local utilities, businesses and science.

Developing a methodology to determine the temporal and spatial performance of acting factors on the affected aquifer system and promoting regional actors to assess and modify its state is the main objective of the bi-national research project ECOAQUA. This is achieved by, among other things, conceptualizing and modeling a local socio-hydrological system and its interactions, as well as capturing and weighting influencing and risk factors.

The contribution of the TU Braunschweig working group to the analysis of the dynamics and complexity of a socio-hydrological consists in applying and extending remote sensing techniques (satellite-based InSAR), cause-specific modeling of surface deformation and data-based descriptive and qualitative hydrological modeling using Support Vector Machine (SVM) algorithms. The latter is facilitated by the high density of relevant field data and remote sensing data. This enables the generation of reliable temporal and spatial information about the state and the transformation processes of the local aquifers, taking into account different data levels.

Here we present preliminary results of this approach.

I. INTRODUCTION

Ensuring availability and sustainable management of water and sanitation for all is one of the great challenges of the 21st century (U.N. General Assembly, 2015). It is well observed that extensive human interaction with the hydrosphere alters the quality and availability of freshwater resources especially in areas with arid to semi-arid climatic conditions. Therefore holistic, efficient and sustainable use of vital resources not only plays a major role on the affected ecosystem, but also for the socio-economic fabric of the area. A guideline set by Vogel et al. in 2015 urged to face this challenge in an integral effort, bringing together knowledge, data and techniques from various disciplines to gain in-depth understanding of the dynamics governing the feedbacks of human and natural systems and approach resolutions of complex water problems.

The bi-national research project ECOAQUA of the Mexican universities UNAM (Mexico City), UABC (Ensenada) and the TU Braunschweig from Germany presents an example of this effort.

The study area is located in the northern part of the Baja California peninsula, an arid costal region 110 km south from the US-Mexican border line. Geologically the region is characterized by coastal and alluvial flatlands, where the city of Ensenada with 280 000 inhabitants and the croplands of Maneadero are found.

Figure 1 *Location of the study area.*

These flatlands are surrounded by the Guadalupe and Ojos Negros intermountain valleys, rising up to 1000 m. The area developed a prosperous economic system with agriculture and livestock breeding being the most important activities, followed by fishing, aquaculture and a rapidly growing touristic sector.

The region depends mainly on water supply from the local aquifers, the coastal Maneadero Aquifer south of Ensenada and the mountainous Guadalupe Aquifer northeast of the city. In the last years growing socio-economic activities, the demand for urban, commercial and tourist infrastructure accompanied by a lack of proper management strategies imposed considerable pressure on the coastal eco- and aquifer system. This has led to considerable shortages in quality and supply of fresh water, saltwater intrusion in the Maneadero Aquifer and prevailing overexploitation of the Guadalupe Aquifer with decreasing water levels and wells running dry, implying adverse effects ranging from soil contamination, crop deterioration and public health indications.

The ECOAQUA interdisciplinary team of researchers from geodesy, geology, mathematics, hydrology, oceanography, sociology, economics, coastal engineering and ecosystem management committed to:

- determine the temporal and spatial performance of acting factors in the area eco- and socio-economic system to enhance its conservation, integration, function and resilience .
- set special focus on groundwater resources and transformation processes of the aquifer system, considering its interdependencies with freshwater supply and soil productivity as basic but crucial services for bio-economic strategies.
- apply and further enhance radar remote sensing data processing techniques in hydrology to determine and weight influencing and risk factors
- establish of a methodology that enables the evaluation and modification of the previous and future performance of the system.

II. ECOAQUA MODELING

A. *From Concept to data-driven Model*

The first but critical step in this research project was conceptualizing a model that would reflect the dynamics of the most crucial components of the coupled natural-human system and help simplify its extreme complexity to a level of understanding its basic functioning. The conceptual model proposed in this study combines the Driver-Pressure-State-Response (DPSIR, Majorosova 2016) with the Source-Pathway-Receptor-Consequence (SPRC, Narayan et al., 2012) framework. The combination of the two, DPSIR being designed to describe human interaction with the environment and SPRC serving as a risk assessment tool, enabled a qualitative description of key components, ergo a holistic description of human and natural forces driving change in the component's states. The basis to select and organize the latter in the DPSIR-SPRC categories strictly followed their relation to adverse effects on the states of the aquifer system and the productivity of the soil, referred to as receptors in the conceptualization. To date 15 adverse states of the receptors with corresponding causality chains (CC) were conceptualized and validated by expert knowledge.

The conceptual model with its component structure was now utilized to explore dynamics and feedback mechanism across scales and tailor data collection and algorithm set-up to perfectly fit the research question. The causality chains guided the design and input data selection for the modeling stage. Our study complements more physically-based approaches in using data-driven methods to integrate multivariate and multitemporal data for setting up prediction models. The family of algorithms used to model the conceptual causality chains are the Support Vector Machines (SVMs). SVM algorithms are based on statistical learning theory and are set up to minimize structural risk, model complexity and prediction error simultaneously (Vapnik, 1999). Their great generalization capability made them attractive for application in the field of hydrological modeling and groundwater level monitoring, taking into account their significant role in managing water resources (Raghavendra & Deka, 2014). Yet this approach is still new and needs to be further exploited to unfold its full potential.

B. *Causality Chain I*

Here we propose an example of such a chain. We study the adverse effect of excessive water extractions to the surface and the vertical compression of soil layers, a phenomenon being referred to as groundwater-induced subsidence. This effect of anthropogenic interaction with the subsurface depends on various site-specific factors such as geology, hydrogeology and geomorphology as well as on pumping rates, land use and population. CCI is an approach to explore this effect and to represent spatial dependencies of the components involved.

The modeling of groundwater-induced subsidence was implemented in Guadalupe Valley, where data availability was sufficient. See Figure 2 for a graphical representation of conceptualized CCI.

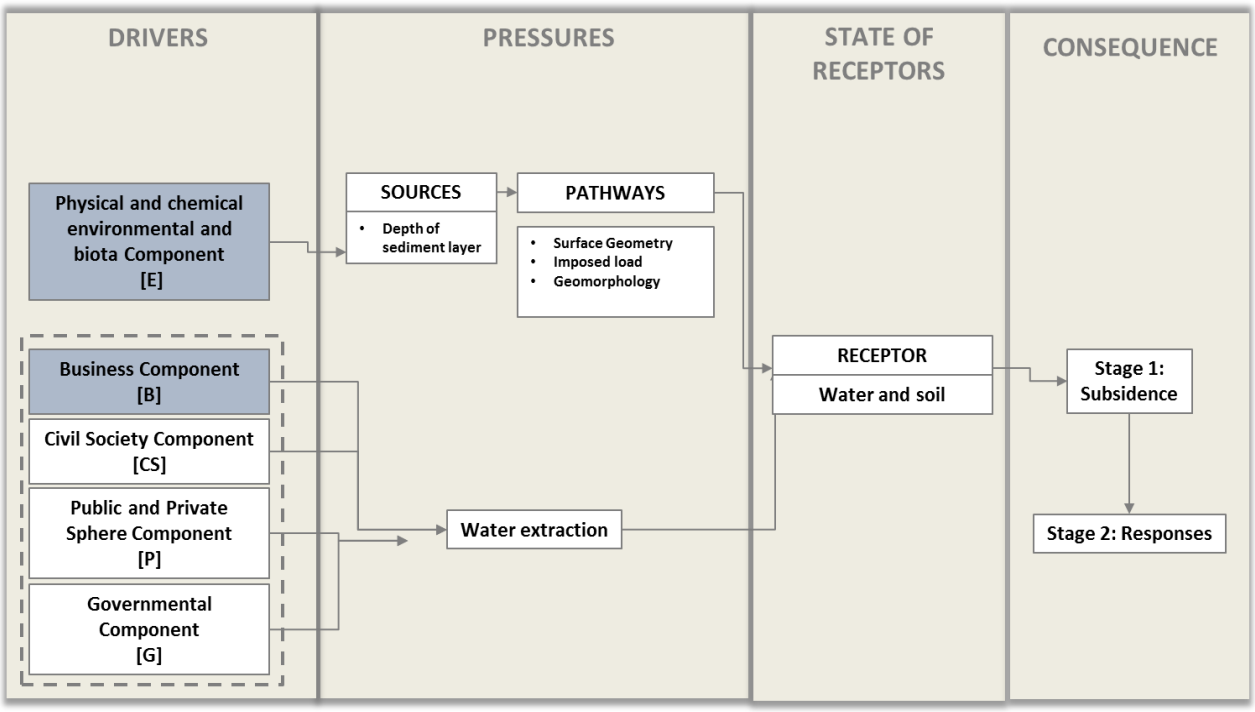

Figure 2 Graphical representation of CCI in the DPSIR- SPRC model. (Carmona & Schottmüller, 2018)

III. CASE STUDY GUADALUPE VALLEY

The Guadalupe Valley is part of the Guadalupe river basin (2380 km²). It is located 38 km NE of the city of Ensenada and covers an area of 333 km². It is formed by a late quaternary basin with even topography and elevations ranging from 280 to 390 meters above sea level. It is surrounded by the Sierra Juarez mountain range (Gastil et al., 1975).

The ephemeral Guadalupe river originates in the Ojos Negros Valley, then flowing west into the Guadalupe Valley. The riverbed traverses the valley and is water bearing during the rainy season from November through March. The alluvial sediment layers produced by this river, consists of gravel, sand, silt and clays in minor proportion.

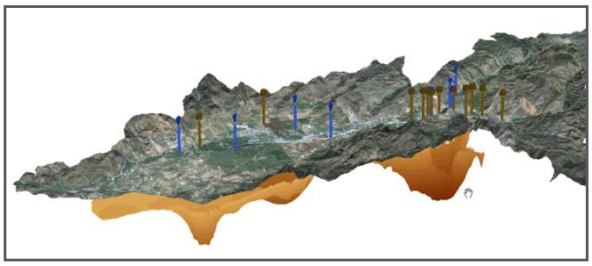

Figure 3: Guadalupe Valley with Guadalupe Aquifer, wells (blue) and lithological drill holes (brown) used in this study.

Guadalupe aquifer constitutes as one of the most important aquifers in the region. Sediment depth of this unconfined aquifer reaches from more than 300 m in the northeast of the valley to 14 m in the south-west (Campos-Gaytan et al., 2014) and extends 79 km². The climate in this area is characterized by large seasonal and annual temperature and precipitation variability with general intense rainfall events and persistent dry periods. Mean annual precipitation averages to 280 mm, mean monthly temperatures range from 0.6°C in winter to 30°C in summer. Groundwater recharge of the catchment area is calculated to reach an annual average of 26.4 Mm³. Since precipitation in the area is deficient to meet crop water requirements, irrigation has become essential in the basin. An estimate of 26.2 Mm³ of groundwater are used for irrigation and local infrastructure and additional 12.4 Mm³ of water is extracted to supply the city of Ensenada, resulting in an annual deficit of 12.2 Mm³(CONAGUA, 2017).

A. *Input data for CCI-SVM regression*

In order to quantify the relation of subsidence to water extractions in the study area, tangible proxies were selected to be fed into the database. Utilizing radar remote sensing data to hydrologic studies promises a monitoring strategy that allows area wide observations in good temporal resolution.

This initial target and feature selection was guided by conceptual knowledge captured in CCI. See table 1 for a summary.

Input Data/ Features			
Features (Proxies)	Source	Description	Proxy for model variable
Elevation	SRTM-DEM, TandemX	Terrain heigth in meters above sea level (masl)	Surface geometry
Slope	SRTM-DEM derived		Surface geometry
Aspect	SRTM-DEM derived	Orientation of slope with respect to direction	Surface geometry
Distance dist(wl, well)	COTAS PNE ABR-2017	Dependent on spatial Resolution of the Label	Superimposed load, water extraction
Distance dist(wl, river)	Optical Data | Sentinel 2	Dependent on spatial Resolution of the Label	Porosity
Depth of sediment layer	Geological Map 1 : 50 000, 11 drill logs, Geoelectrical survey	Depth in meters below top ground surface	Infiltration capacity
Output Variable/ Target			
InSAR velocites	Sentinel 1	Annual vertical velocity	subsidence

Table 1 Description of initial training data for the SVM – modeling of Causality Chain I

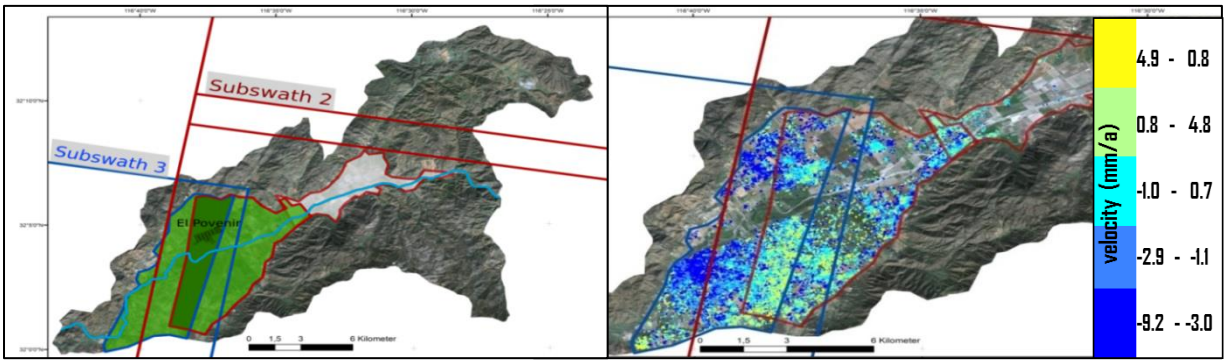

Figure 4 **A** Study area, Sentinel-1A subswaths in red and blue and the extracted areas for advanced processing within one subswath in green (SBAS) and grey (PSI).
B Results from SBAS and PSI for Guadalupe Valley from October 2014 until June 2017.

B. *InSAR – derived vertical velocity*

The radar data to derive vertical velocities was captured by the Sentinel-1A satellite. The Guadalupe Valley is represented in descending track 173 and contains 47 partially overlapping acquisitions from October 2014 to June 2017, see figure 4.

To derive time-dependent deformation velocities two multi-temporal approaches, the Persistent Scatterer Interferometry (PSI)[1] and the Small Baseline Subset (SBAS)[2] algorithm were adopted to process the data. This dual approach is justified by the characteristics of the study area. The SBAS technique leads to stable results especially in rural areas, whereas PSI performs best in areas with temporally coherent point targets such on houses and infrastructures.

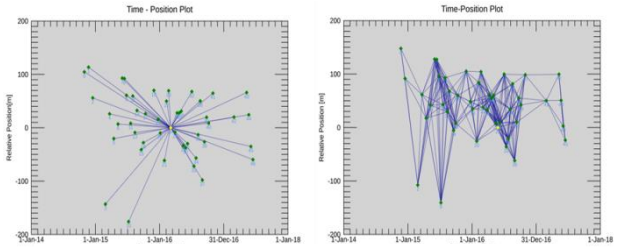

Figure 5 Connection graphs for PSI and SBAS processing.

Figure 5 depicts the spatio-temporal connections between the radar scenes for the PSI and the SBAS processing. Whereas for PSI processing one master scene is chosen to produce interferograms (left) with the remaining scenes (slaves), the SBAS technique utilizes different master and slave scenes, chosen according to a maximum difference in data of acquisition. A total of 270 interferograms entered the respective processing routines. In Figure 4B the joined solution from the preliminary results of the InSAR processing are illustrated.

Figure 4A depicts the study area and the available InSAR subswath ensemble. In this first processing approach three independent solutions with velocity information were generated. In the eastern part spatial overlap of the solutions is attributed to the different processing techniques, whereas in the

[1] developed by Ferretti et al. (2001)
[2] developed by Berardino et al., (2002)

western part around El Porvenir it was produced by the overlap of Subswath 2 and 3.

Out of the 9 836 pixels with two different velocity values, the one with higher coherence was chosen to enter the final data set. After outlier removal a total of 76 476 values for the target variable in the first SVM causality chain where prepared.

IV. MACHINE LEARNING-MODEL AND PRELIMINARY RESULTS

A. *ε–Support Vector Regression and hyper-parameters*

The specific machine learning algorithm used in this study is the ε-Support Vector Regression (ε-SVR). Here we provide only basic ideas behind ε-SVR and relevant model parameters for implementation. For comprehensive information on statistical learning theory and SVMs please refer to Vapnik (1999), Smola & Schölkopf (2004) and Burges (1998).

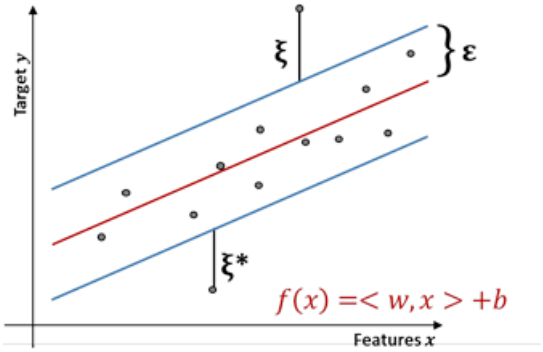

Figure 6 Example of 1D linear regression with ε – insensitive zone

Given a set $(\boldsymbol{x_1}, y_1), \ldots, (\boldsymbol{x_n}, y_n) \in (X \times \mathbb{R})$ of observations, the aim of ε-SVR is to find a function f, such that $f(\boldsymbol{x_i})$ deviates ε at most from the observed target y_i for all training examples and is at the same time as flat as possible to reduce model complexity. To allow for some errors greater than ε, additional slack variables ξ, ξ* measuring the deviations greater than ε are introduced. The slack variables represent additional constrains to the system output with their impact being regularized by the cost parameter C.

The ε – SVR optimization task is not dependent on the dimensionality of the feature space, but only on dot product of the data in it. When trying to reveal non-linear relationships with more complex functions, we can use the so-called kernel trick. Employing this trick means performing linear regression in a space of higher dimension, using the kernel to transform the dot product of our data in that space. In this study we used the Gaussian kernel k to perform the mapping. It is given by

$$k(x, y) = exp(-\gamma \parallel x - y \parallel^2) \qquad (1).$$

It is well known that SVM model complexity and hence generalization performance highly depends on choice of the model and the kernel parameters ε, C and γ (see for example Vapnik, 1999). A general description of their influence is:

- **ε** defines the width of the insensitive zone and therefore affects the number of support vectors used to construct the regression function.
- **C** controls the trade-off between flatness and the degree to which deviations greater than ε are tolerated.
- **γ** is the inverse of the standard deviation of the Gaussian kernel. It defines the influence of a single training example on the construction of the decision boundary, ergo the choice of support vectors in the training set.

B. *Spatial Autocorrelation*

Spatial autocorrelation of training data poses a severe violation of the model assumptions on the independence and identical distribution of the observations. This can affect prediction accuracy substantially and lead to overoptimistic prediction results (Brenning, 2005).

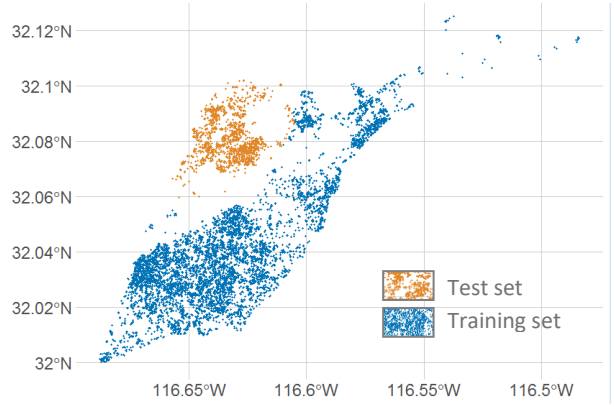

Figure 8 Example spatial partitions in Guadalupe Valley.

Spatial autocorrelation occurs, because points geographically close to each other are in general more similar than those further away. Therefore a partition strategy that splits the data into spatially disjoint train and test subsets was implemented into model building and prediction. Here we utilized the distance-based spatial partitioning as implemented in R, using the observations' coordinates in a *k*-means clustering.

C. *Feature Selection and Tuning*

In this study a stepped optimization of the input feature subset and the optimal performing hyper-parameters was utilized, taking into account the interdependency of the latter. The first step incorporated fast and effective hyper- parameter selection directly from the training data as proposed by Cherkassky & Ma (2002). The width of the ε-

insensitive zone was chosen with respect to expected noise in the radar velocities. Then feature selection was carried out with different filter and wrapper methods, its performance evaluated and validated on the specific SVR model. Whereas all filter methods indicated the slope variable to be redundant, the computational expensive wrapper method Sequential Backward Selection (SBS) repeatedly removed all geometric input features, namely elevation, slope and aspect to substantially decrease the risk over-fitting. On the SBS- optimized feature subset the last optimization step was to fine tune hyper-parameters on the reduced feature with a 5-fold spatial cross-validated random search.

V. RESULTS

A. *Structure of CCI steady state SVM- model*

Spatial modeling task: Regression task with target velocity and 6 features
Learner algorithm: ε - SVR with Gaussian kernel
Hyper-parameters: ε, C , γ
Performance measure: RMSE
Performance level for stepwise spatial tuning of hyper-parameters and input feature set:
5-fold cross-validation with 5 repetitions

B. *Final model and results*

The final prediction model run for the target *vertical velocity* with 3 input features {*dist_well, dist_river, sediment_depth}* and hyper-parameter set (ε, C, γ) =(2, 5, 0.005) resulted in deviations between observed and predicted in the order of 2.92mm root mean squared error (RMSE). This gap can be regarded as rather large with respect to the observed target. To diagnose on causes for those deviations, we utilize the learning curve.

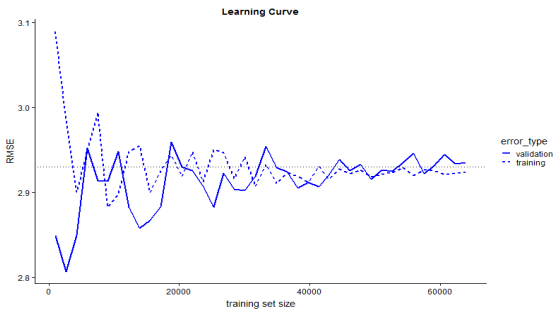

Figure 8 Learning curve of the ε – SVR with optimal input feature set and hyper-parameters.

This curve gives fast and valuable insight into the generalization performance of the learned relationship of target and features. It displays the relation between the number of training examples and the error scores on the training and validation set. Thereby it helps diagnosing bias and variance to reduce errors in the algorithms set-up. From Figure 8 we can derive that the learner performance reaches stability for a training set size of about 45000. The curve displays a good fit to the data, still its peaky progression indicates that train and test set splits need to be harmonized in terms of statistical properties. Finally we observe that generalization performance is most likely to suffer from the small set of input features with explanatory variables missing.

VI. CONCLUSION

The conceptualization in the DPSIR-SPRC framework constitutes a theoretically sound basis for the combination of multiple layers of data in the modeling stage. This was of particular importance in the case of those layers that are either not considered, or modeled as external factors in numerical or analytical physically - based models. Especially the role of human activities in the hydrologic cycle with its many facets is an instance of such data.

After a number of repetitive tuning steps to increase model performance, it became clear that the data set itself introduced a great part of the deficiency. To a certain extent this was expected since proxies for CCI model variables were poorly available at the time when data acquisition for this study was completed. This has now changed, especially for the target variable. An enhanced velocity time-series processing has meanwhile been completed, utilizing Sentinel 1 data from 2014 until February 2019. Taking the special characteristics of the study area into account, a coherent, geology oriented solution was generated by a consistent SBAS-processing. The feature with highest influence on the target, the distance to wells being proxy for water extractions, will be replaced by actual observations of water levels from piezometers. Those changes will supposedly lead to a great impact on the generalization performance of the regression model.

ACKNOWLEDGEMENTS

This research is funded by the Federal Ministry of Education and Research (BMBF) with Research Grant ID 01DN16035. The Sentinel-1 and -2 data are provided by Open Access Copernicus, European Space Agency (ESA).

References

Berardino, P., G. F., Lanari, R., & Sansosti, E. (2002). A new algorithm for surface deformation monitoring based on Small Baseline differential SAR Interferometry. IEEE Trans. Geosc. Rem. Sens, Vol. 40, No. 11.

Brenning, A. (2005). Spatial prediction models for landslide hazards: review, comparison and evaluation. Natural Hazards and Earth System Sciences, 5, 853-862.

Burges, C. J. (1998). A Tutorial on Support Vector Machines for Pattern Recognition. Data Mining and Knowledge Discovery, 2, 121-167.

Campos-Gaytan, J. R., Kretzschmar, T., & Herrera-Oliva, C. S. (2014). Future groundwater extraction scenarios for an aquifer in a semiarid environment: case study of Guadalupe Valley Aquifer, Baja California, Northwest

Mexico. Environmental Monitoring and Assessment, 186, 7961-7985.

Carmona, R. E., & Schottmüller, C. (2017). Application of a Source-Pathway-Receptor- Consequence (SPRC) conceptual model to overexploited aquifer system in a coastal arid region. Tech. rep., Universidad Nacional Autonoma de Mexico (UNAM).

Cherkassky, V. &. (2002). Selection of Meta-Parameters for Support Vector Regression. ICANN 2002, LNCS 2415, pp. 687–693.

CONAGUA. (2017). Actualización de la dispoibilidad media anual de agua en ael acuífero Guadalupe Banuelos. Diario Oficial de la Federación.

Ferretti, A. a., & Rocca, F. (2001). Permanent scatterers in SAR interferometry. Geoscience and Remote Sensing, IEEE Transactions on, vol. 39, no. 1.

Gastil, R. G., Phillips, R. P., & Allison, E. C. (1975). Reconnaissance geology of the State of Baja California (Memoir 140- Geological Society of America). Geological Society of America.

Majorosova, M. (2016). DPSIR framework - A decision-making tool for municipalities. Slovak Journal of Civil Engineering, 24(4), pp. 45-50.

Narayan, S. a., & Monbaliu, J. (2012). A holistic model for coastal flooding using system diagrams and the Source-Pathway-Receptor (SPR) concept. Natural Hazards and Earth System Sciences.

Raghavendra, N. S., & Deka, P. C. (2014). Support vector machine applications in the field of hydrology: A review. Applied Soft Computing, 19, 372-386.

Smola, A. J., & Schölkopf, B. (2004). A tutorial on support vector regression. Statistics and Computing, 14, 199-222.

Vapnik, V. N. (September 1999). An Overview of Statistical Learning Theory. IEEE Transactions on Neural Networks, Vol. 10, p. 988 -999.

Vogel, R. M., Lall, U., Cai, X., Rajagopalan, B., Weiskel, P. K., Hooper, R. P., et al. (2015). Hydrology: The interdisciplinary science of water. Water Resources Research, 51, 4409-4430.