ORIGINAL PAPER

Combination of geostatistics and self-organizing maps for the spatial analysis of groundwater level variations in complex hydrogeological systems

Emmanouil A. Varouchakis[1] · Dimitri Solomatine[2,3] · Gerald A. Corzo Perez[2] · Seifeddine Jomaa[4] ·
George P. Karatzas[5]

Abstract

Successful modelling of the groundwater level variations in hydrogeological systems in complex formations considerably depends on spatial and temporal data availability and knowledge of the boundary conditions. Geostatistics plays an important role in model-related data analysis and preparation, but has specific limitations when the aquifer system is inhomogeneous. This study combines geostatistics with machine learning approaches to solve problems in complex aquifer systems. Herein, the emphasis is given to cases where the available dataset is large and randomly distributed in the different aquifer types of the hydrogeological system. Self-Organizing Maps can be applied to identify locally similar input data, to substitute the usually uncertain correlation length of the variogram model that estimates the correlated neighborhood, and then by means of Transgaussian Kriging to estimate the bias corrected spatial distribution of groundwater level. The proposed methodology was tested on a large dataset of groundwater level data in a complex hydrogeological area. The obtained results have shown a significant improvement compared to the ones obtained by classical geostatistical approaches.

Keywords Transgaussian Kriging · Geostatistics · Self-organizing maps · Machine learning · Groundwater · Box-Cox

1 Introduction

Water scarcity is a major global problem, and it is expected to increase significantly in the near future. Projections for the Mediterranean region indicate a gradual decline in runoff and significant changes in surface and groundwater availability (García-Ruiz et al. 2011; Hartmann et al. 2014). The importance of groundwater in mitigating and buffering the problems associated with water scarcity in the Mediterranean region has been clearly recognized, especially in light of the increasing impacts of climate change (Jomaa Seifeddine et al. 2021; Smerdon 2017; Thomas and Famiglietti 2019). The development of adequate modeling tools is recommended in informing precautionary water resources management plans and allowing adaptation to the projected changes (Garrote et al. 2015; Wada et al. 2010).

Large scale groundwater modelling is complex due to the various boundary conditions and the hydrogeological inhomogeneities that usually exist. Geostatistical methods have often been used as a surrogate approach to model the groundwater level spatial variability. However, in complex groundwater systems with inhomogeneous aquifers, geostatistical analysis of monitored groundwater levels is challenging because stationarity is not guaranteed, and

✉ Emmanouil A. Varouchakis
  evarouchakis@tuc.gr

1 School of Mineral Resources Engineering, Technical University of Crete, Chania, Greece

2 IHE Delft Institute for Water Education, 2601 DA Delft, the Netherlands

3 Delft University of Technology, Delft, the Netherlands

4 Department of Aquatic Ecosystem Analysis and Management, Helmholtz Centre for Environment Research - UFZ, Magdeburg, Germany

5 School of Chemical and Environmental Engineering, Technical University of Crete, Chania, Greece

 Springer

spatial dependence of observations cannot be reliably estimated. Furthermore, in the case of areas with rich monitoring networks, the availability of a large dataset of observations requires an efficient preprocessing algorithm to organize the dataset based on its characteristics.

In this work, a geostatistical analysis approach by means of Ordinary Kriging (OK) in combination with a machine learning algorithm, Self-Organizing Maps (SOM), was applied (a) to obtain stationary groundwater level data for producing a more stable variogram that reaches a plateau (Goovaerts 1997), leading to efficient estimation of observations spatial dependence (b) to develop reliable spatial maps of groundwater level variability in a large-scale hydrogeological system of complex aquifers, and (c) to process a large dataset of 2524 wells efficiently using geostatistical analysis. The proposed approach represents an innovative approach in geostatistics by using the original SOM technique to classify similar neighboring data to create stationary neighborhoods and perform geostatistical analysis employing the OK method. OK and SOM are combined into a GeostatSOM hybrid approach.

Geostatistics has been successfully applied in groundwater-related studies (Cushman and Tartakovsky 2016; Kitanidis 1997; Panagiotou et al. 2022), and in combination with data science techniques to improve groundwater level spatial variability estimation (mapping). An indicative list of such combined research works follows: Hoogland et al. (2010) applied the multiple linear regression technique in terms of areal data to interpolate by means of Regression Kriging (RK) groundwater depths with improved accuracy, Manzione and Castrignanò (2019) employed multivariate data fusion methodologies along with OK to predict groundwater levels, Theodoridou et al. (2017) perform spatial analysis of groundwater levels using OK employing a Fuzzy Logic approach, Varouchakis and Hristopulos (2019) applied an exponential weighted moving average process to detrend groundwater level observations and improve its spatial analysis utilizing RK, and Varouchakis et al. (2019) interpolated groundwater levels using a Bayesian kriging approach for advanced accuracy.

Several studies have shown that the SOM technique can be successfully applied to water resources topics e.g., in clustering and classifying large datasets of precipitation, runoff and drought to support watershed and large-scale hydrologic modelling approaches (Kalteh et al. 2008; Markonis et al. 2018), organising large collections of hydrometeorological data for streamflow and precipitation prediction (Hsu and Li 2010; Toth 2009) and rainfall-runoff modelling using satellite products by classifying satellite imagery data (Farzad and El-Shafie 2017; Nourani et al. 2013).

SOM can be successfully combined with geostatistics in spatial analysis applications to classify the properties of large datasets (Kanevski 2013); it has been successfully applied in soil science to express the nonlinear relationships between soil organic matter and correlated factors in terms of regression kriging (Huang et al. 2017). However, to our knowledge, this is the first time such a hybrid approach has been applied to groundwater level spatial analysis. A similar methodological approach called Geo-SOM (Henriques et al. 2012) considers the coordinates of available observations to perform data clustering only.

The island of Crete in Greece is used as a case study due to its rich groundwater level monitoring network and complex hydrogeological system to verify the proposed methodology. The island's groundwater resources have been significantly affected by overexploitation and are expected to be further affected by the climate change projections for the Eastern Mediterranean (Founda et al. 2019; Kim et al. 2019; Koutroulis et al. 2013; Special water secretariat of Greece 2017).

This work could be a guide for analogous applications in areas with complex hydrogeologic systems and large data sets where physically based models require a significant amount of information and conventional geostatistical approaches are not directly applicable. Thus, scientists and policymakers can apply it to determine and understand the state of the aquifer and its variability in complex systems identifying areas where groundwater management interventions are needed.

Apart from the introduction section, where background information on the proposed methodologies and its relativeness to the topic are presented, the remainder of the paper is formed in three sections. In materials and methods section the case study and data availability is discussed, and the mathematical background of the proposed methods and its implementation is presented. The results and discussion section follows, where the proposed research findings are described and discussed, and finally, in the last section the concluding remarks are highlighted.

2 Materials and methods

2.1 Study area and data availability

The island of Crete is located in the southeastern part of the Mediterranean basin (Fig. 1). Hydrogeologically, it consists of a mixed sequence of porous and karstic aquifers. The most important hydrogeologic systems in terms of capacity and supply capabilities are the karst aquifers formed in carbonate rocks. These are followed by alluvial aquifers which are generally intensively exploited due to their relatively shallow depth. Finally, low groundwater capacities are found in rocks such as phyllites, schists, ophiolites which are important for water consumption as

Fig. 1 Location of the island of Crete, Greece. Aquifers' system distribution in the island; location of the corresponding 2524 monitoring wells employed in the present study
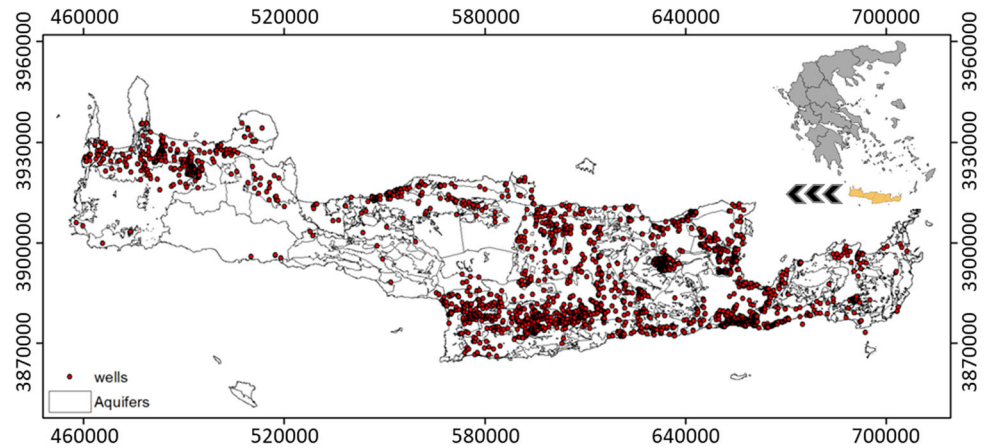
they meet the water needs of high-altitude settlements. The hydrogeological units of Crete have an area of 6.303 km^2 compared to the total area of Crete which is 8.335 km^2. The remaining area of about 2000 km^2 relates mainly to outcrops of phyllite—quartzite and flysch and small occurrences of carbonate and granular formations (Neogene–Quaternary) which are not related to the main hydrogeological units of the water district (Special water secretariat of Greece 2017).

In the 1980s, an extensive network of pumping stations was installed in Crete and dry cultivation was converted to drip-irrigated. As a result, productivity has increased, but this overuse has led to a dramatic decline in groundwater levels in many aquifers over the past thirty years. In the water district of Crete, although a number of reservoirs have been built recently, the main source of irrigation has always been groundwater (drilling of natural sources) (Croke et al. 2000; Special water secretariat of Greece 2017; Varouchakis 2016).

In the most recent Water Resources Management Plan prepared under the Water Framework Directive for the Crete Water District, the quantitative and qualitative status of the island's groundwater bodies was assessed in addition to the surface water catchments. Out of the 91 groundwater bodies, 9 are of poor-quality, and 10 are significantly depleted. Due to their qualitative and quantitative characteristics, 11 groundwater bodies require special attention (Special water secretariat of Greece 2017). In addition, two recent studies confirmed the findings by assessing the recharge potential of these groundwater bodies (Matiatos and Wassenaar 2019; Varouchakis et al. 2018).

In this work, the monitored biannual average data of 2524 wells (Special water secretariat of Greece 2017, 2020) are processed for the hydrological year 2017/18 (Table 1), using the GeostatSOM approach to estimate the spatial variability of the water table in a complex hydrogeological system of aquifers (Fig. 1).

2.2 Spatial statistics

The classical interpolation method Ordinary Kriging (OK) (Cressie 1990; Krige 1951; Matheron 1971) was used in this work to combine SOM with geostatistics. Kriging provides an estimate of the variable and the error variance of the corresponding estimate (the associated uncertainty). In ordinary kriging, the error variance (1) depends on the variogram model (spatial interdependence); the estimation accuracy depends on the complexity of the spatial variability of the random field as modeled by the variogram, (2) depends on the data configuration and its distances from the location being estimated, (3) is independent of the data values; for a given variogram model, two identical data configurations give the same variance regardless of their values and (4) the error variance is zero at the data locations and increases in distance from the data while it reaches a maximum value in the extrapolation situation (Goovaerts 1997; Kitanidis 1997).

However, for highly skewed data, a Gaussian transformation approach is required for Kriging error variance to be reliable as it is possible to create confidence intervals (Clark and Harper 2000; Dowd 2018). Kriging maps variability only up to the second order moment (covariance); therefore, the random field of the transformed variables must be Gaussian to derive unbiased estimates at unsampled locations (Deutsch and Journel 1992; Kitanidis 1997; Varouchakis 2021). OK is known to be optimal when the data have a multivariate normal distribution, and the true variogram is known. Therefore, data transformation is required before kriging to normalize the data distribution, suppress outliers, and thus improve variogram estimation (Armstrong 1998; Deutsch and Journel 1992). Estimation is performed in the Gaussian domain; before the estimates are transformed back to the original domain. However, the back transformation of the data may lead to biased results (Belitz and Stackelberg 2021; Schabenberger and Gotway 2005). Therefore, in this work the Transgaussian Kriging

Table 1 Statistical measures of the groundwater level data. z_{min}: minimum value; $z_{0.25}$ first quartile; $z_{0.50}$ median; m_z mean; $z_{0.75}$ third quartile; z_{max} maximum value; $\hat{\sigma}_z$ standard deviation; $\hat{s}_z$ skewness coefficient; $\hat{k}_z$ kurtosis coefficient

z_{min}	$z_{0.25}$	$z_{0.50}$	m_z	$z_{0.75}$	z_{max}	$\hat{\sigma}_z$	$\hat{s}_z$	$\hat{k}_z$
0.2 masl	10.00 masl	15.29 masl	36.80 masl	58.20 masl	789 masl	67.43 masl	3.64	19.34

method in terms of the OK predictor is applied that corrects the retransformation bias (Cressie 1993).

Transgaussian Kriging (TGK) is recommended for geostatistical analysis of data that require a Gaussian transformation (Schabenberger and Gotway 2005) and has been used successfully with groundwater data (Varouchakis et al. 2012). It employs the Box-Cox transformation method. For a given data set of measurement values Z at corresponding measurement locations $\mathbf{s}$, $Z(\mathbf{s})$, a nonlinear normalizing transformation is expressed as $Y(\mathbf{s}) = g(Z(\mathbf{s}))$, where $Y(\mathbf{s})$ follows the multivariate Gaussian distribution. Implicitly, $Z(\mathbf{s}) = \varphi(Y(\mathbf{s}))$, where $\varphi(\cdot) = g^{-1}(\cdot)$ is a one-to-one, twice-differentiable function. It is also assumed that $Y(\mathbf{s})$ is an intrinsically stationary random field with mean m_Y and variogram $\gamma_Y(\mathbf{r})$. For an unknown m_Y, the predictor OK, $\hat{Y}_{OK}(s_0)$, is used to predict $Y(\mathbf{s}_0)$. An estimate of $Z(\mathbf{s}_0)$ is then given by $\hat{Z}(\mathbf{s}_0) = \varphi(\hat{Y}_{OK}(\mathbf{s}_0))$, where $\varphi(\cdot)$ is the inverse of the transformation function. However, this leads to a biased predictor if $\varphi(\cdot)$ is a nonlinear transformation. A bias-correcting approximation is the trans-Gaussian predictor (Cressie 1993):

$$\hat{Z}(\mathbf{s}_0) = \varphi(\hat{Y}_{OK}(\mathbf{s}_0)) + \frac{\varphi''(\hat{m}_Y)}{2}\left[\sigma^2_{OK;Y}(\mathbf{s}_0) - 2\mu_Y\right] \qquad (1)$$

where $\hat{m}_Y$ is the OK estimate of m_Y, μ_Y is the Lagrange multiplier of the OK system, $\varphi''(\cdot)$ is the second-order derivative of the inverse transformation function, and $\sigma^2_{OK;Y}(\mathbf{s}_0)$ is the OK variance. When the Box-Cox normalizing transformation is used, as here, the functions $\varphi(\cdot)$ and $\varphi''(\cdot)$ have the following form (Cressie 1993):

$$\varphi(Y) = (\lambda Y + 1)^{1/\lambda} \qquad (2)$$

$$\varphi''(\hat{m}_Y) = (1 - \lambda)\,(\hat{m}_Y \lambda + 1)^{\frac{1}{\lambda}-2}. \qquad (3)$$

where λ is the power exponent of the Box-Cox method (Box and Cox 1964).

The Spartan variogram function was used because it has been successfully applied in spatial and space–time geostatistical data analysis of groundwater level data (Ruybal et al. 2019; Varouchakis and Hristopulos 2013, 2019). To evaluate the efficiency of the proposed method it was compared with the well-known Regression Kriging method in terms of auxiliary variables such as coordinates and elevation, Universal Kriging and Ordinary Kriging.

Statistical metrics such as the Mean Absolute Relative Error, the Mean Absolute Error and the Correlation Coefficient (R) were employed to evaluate the comparison of the methods.

2.3 Self-organizing maps

Self-organizing maps (or Kohonen networks) are an unsupervised machine learning technique (Kohonen 2001b). It is used for dimensionality reduction of data, clustering of complex datasets, local data similarity detection and pattern identification (Bowden et al. 2005; Hsu and Li 2010; Kohonen 2013; Richardson et al. 2003). SOM transforms complex statistically based relationships of high-dimensional data into simpler spatial or temporal relationships that can be represented as an easily interpretable map of a lower dimension. In a sense, SOM can be viewed as a clustering technique, that leads to the identification of data groupings (clusters), i.e., measurements (vectors) in the training set. However, it is different from the other clustering methods, which result in assigning each measurement location to one of the clusters, as is the case, for example, with the widely used k-means clustering algorithm. SOM yields a set of nodes (typically arranged in the form of a rectangular or hexagonal two-dimensional grid) whose positions (in terms of Euclidean distance) would be close to groups of measurement values. A trained SOM can then be used as a classifier: it classifies a vector from the input space by finding the node (called the best matching unit, BMU) that is closest to that input vector. Initially, the SOM nodes, forming a grid, are randomly located and gradually move towards the groups of data measurements during the iterative training process.

The algorithm below (Kohonen 2001a) presents the basic steps of training the SOM network consisting of a certain number of output nodes. A data set $Z(\mathbf{s})$ of n locations, $\{s_1, \ldots s_n\}$ in k-dimensional space, is considered:

Self-organizing maps learn to classify input vectors according to how they are grouped in the input space. They learn to recognize neighboring sections of the input space. Thus, self-organizing maps learn both the distribution and topology of the input vectors they are trained on. The nodes in the layer of a SOM are arranged originally in physical positions according to a topology function, such as a grid or hexagonal in such a way that topological properties in the

input training data are preserved (Augustijn and Zurita-Milla 2013; Kohonen 2013).

We used the Kohonen R package (Wehrens and Buydens 2007) to train several SOMs following these steps:

1. The size (number of nodes, including number of rows and columns) and type (rectangular) of the SOM lattice were chosen.
2. Each node was assigned a random vector of weights ($\mathbf{m}_k$) with the same dimensionality as the input data.
3. Random data samples (input vector) were iteratively presented to the low-dimensional lattice to identify the best matching unit, BMU, which is the node that minimizes the Euclidean distance between the weight vector and the data sample at hand. This iterative process is known as training the SOM and each iteration (t) is used to update the weight vector of the BMU and the neighbouring nodes according to:

$$\mathbf{m}_k(t+1) = \mathbf{m}_k(t) + \alpha(t)(\mathbf{x}(t) - \mathbf{m}_k(t)) \qquad (4)$$

in which $\mathbf{m}_k$ is a k-dimensional weights vector, $\alpha(t)$ is the learning rate and $\mathbf{x} \in Z_k(\mathbf{s})$ is a randomly chosen input vector from the training dataset (Augustijn and Zurita-Milla 2013; Kohonen 2013; Wehrens and Buydens 2007). The aforementioned methodological steps are summarized in a flowchart (Fig. 2)

Thus, when a vector $\mathbf{x}$ is presented, the weights of the winning node and its close neighbors move towards $\mathbf{x}$. The winning node is rewarded with becoming more like the sample vector. Consequently, after many iterations, neighboring nodes have learned vectors similar to each other. For SOM training, the weight vector associated with each node moves to become the center of a cluster of input vectors. After training the SOM, the next step in the SOM process is mapping the data onto the trained SOM network, identifying for each node location in the topology how many of the observations are associated with each node (cluster centers). This process is repeated for each input vector for a (usually large) number of cycles. The network winds up associating nodes with the input data set. This is delivered by the SOM Sample Hits plot (Kohonen 2013; Wehrens and Buydens 2007).

The quality and accuracy of SOM results can be determined by two performance indices, quantization error, and topographic error. The quantization error is the average distance between the data vector and the best matching unit (BMU) (the closest node). The quantization error indicates how accurately the given input patterns are represented by a BMU. The smaller this index, the better the representation. The topographic error is related to the quality of the map topology: the quality is high (and thus the error low) if the nodes adjacent on the grid correspond to similar input patterns (Beale and Jackson 1990). It is calculated as the
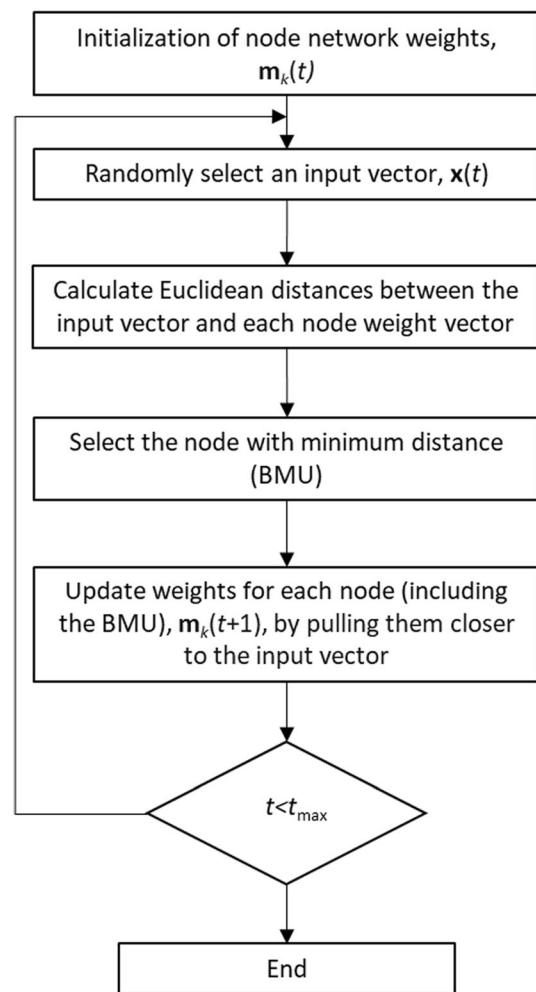
Fig. 2 Flowchart of the unsupervised SOM algorithm (Kohonen 2001a) based on Eq. (4); $\mathbf{m}_k$ is a k-dimensional weights vector, $\mathbf{x}$ is an input vector from the dataset, BMU is the Best Matching Unit, t the iterations and t_{max} the maximum iterations number

ratio of all vectors, where the first and second best BMUs are not adjacent on the grid.

2.4 Implementation and analysis

In this study the two-dimensional rectangular SOM was used. The SOM method was implemented in R using the related package (Wehrens and Buydens 2007) by the authors, while the code originally developed in R which combines Transgaussian kriging with SOM (TGK-SOM), and forms the GeostatSOM model, was developed using the geoR package (Ribeiro et al. 2007). This package was also used to perform Regression Kriging, Universal Kriging and Ordinary Kriging. The cross-validation measures used to compare the true and estimated groundwater level values using the different methods are:

Mean absolute error (MAE):

$$\varepsilon_{\mathrm{MA}} = \frac{1}{N}\sum_{i=1}^{N}\left|z^*(\mathbf{s}_i) - z(\mathbf{s}_i)\right| \qquad (5)$$

Mean absolute relative error (MARE):

$$\varepsilon_{\mathrm{MAR}} = \frac{1}{N}\sum_{i=1}^{N}\left|\frac{z^*(\mathbf{s}_i) - z(\mathbf{s}_i)}{z(\mathbf{s}_i)}\right| \qquad (6)$$

Linear Correlation Coefficient:

$$R = \frac{\sum_{i=1}^{N}\left[z(\mathbf{s}_i) - \overline{z(\mathbf{s}_i)}\right]\left[z^*(\mathbf{s}_i) - \overline{z^*(\mathbf{s}_i)}\right]}{\sqrt{\sum_{i=1}^{N}\left[z(\mathbf{s}_i) - \overline{z(\mathbf{s}_i)}\right]^2}\sqrt{\sum_{i=1}^{N}\left[z^*(\mathbf{s}_i) - \overline{z^*(\mathbf{s}_i)}\right]^2}} \qquad (7)$$

where $z \in Z$; $z^*(\mathbf{s}_i)$ is the estimated level at location $\mathbf{s}_i$, $z(\mathbf{s}_i)$ the observed value, $\overline{z(\mathbf{s}_i)}$ denotes the spatial average of the data and $\overline{z^*(\mathbf{s}_i)}$ the spatial average value of the estimates.

The SOM method was used in this work to help identify the kriging estimation neighborhood. As is well known, the correlation length parameter calculated from the variogram estimation is used to define the estimation neighborhood. However, in many cases the theoretical variogram model calculation is the best possible without perfectly modeling the experimental variogram, which is based on the measurements and usually does not have a clear sill, which is an important factor in determining a representative correlation length. This is usually the case when long-distance spatial variations exist, and measurement values with high variability lead to significant fluctuations in variogram values. Another reason for the experimental variogram not reaching a sill is the presence of significant trends. Therefore, the calculation of the correlation length is uncertain.

To obtain an appropriate correlation length, various methods are used, such as detrending the data or cross validation using variable correlation lengths based on the rule that a tolerance can be applied. In addition, simulation methods such as sequential Gaussian simulation are also applied, where different realizations are developed that also take into account the uncertainty of the variogram parameters. At the same time, the variability map of the measurements usually depends on the mean of the simulations.

Thus, in this work, SOM was applied to identify the locally similar input data measurements by using the associate Kohenen network node associated with that local neighborhood. The measurement locations belonging to the radius of influence of a node have as their estimation neighborhood the measurements that are associated with that node.

Another reason for using SOM to identify the estimation neighborhood is the nature of the case study considered, the island of Crete. The spatial variability of the groundwater level is studied in a large-scale complex hydrogeological system of porous and karst type aquifers. The physically based modelling of the whole system requires prior knowledge of all physical and anthropogenic boundaries and factors affecting the groundwater level variations. On the other hand, a geostatistical approach may seem more convective as a surrogate, since only the spatial/spatiotemporal dependence needs to be determined. However, in a large-scale application with the presence of a sequence of different aquifer types the obtained correlation length may not be representative of the aquifer type. Our experience shows that in some similar cases, the obtained correlation length is long and considers values of adjacent aquifers of different types, which negatively affects the estimation procedure. To avoid similar problems, this study applies the SOM method to cluster the data based on local similarity, and we consider this an innovative addition to the geostatistics methodology.

3 Results and discussion

The original groundwater level dataset is significantly skewed (Table 1). Therefore, the Box-Cox transformation method, with calculated power exponent $\lambda = -0.32$, was employed to transform the data closer to the Gaussian distribution. The data Gaussianity was significantly improved (Fig. 3), and this is also validated by the improvement in the metrics: $\hat{s}_Y = 0.25$, $\hat{k}_Y = 3.00$, very close to the desirable values: $s = 0$, $k = 3$.

The global experimental variogram of the entire observations' dataset (Fig. 3) exhibits significant fluctuations and does not appear to level off (reach a sill). On the other hand, the one of the transformed data set (Fig. 4) is more stable, and exhibits less fluctuations without reaching a clear sill as well. The best fit theoretical variogram model (Fig. 4) was determined after testing on the experimental variogram different lag distances and maximum variogram spatial distance. The main reason for these variations is the considerable inhomogeneities of the aquifer systems, which significantly affect the piezometric values of the groundwater and lead to significant variations in the level range.

A 19-by-5 nodes grid was applied in this work (Fig. 5) due to the dimensions of the island of Crete and 10,000 repetitions to determine the most representative SOM. When the data measurements associated with each node (Fig. 6) are sufficient to develop at least 30 pairs for each experimental variogram lag distance (Deutsch and Journel 1992), a new variogram is calculated. Then the calculated local parameters (after theoretical variogram fitting) are applied in the TGK prediction procedure for the estimation locations under the influence of the node, if not the

Fig. 3 Experimental variogram of the original dataset
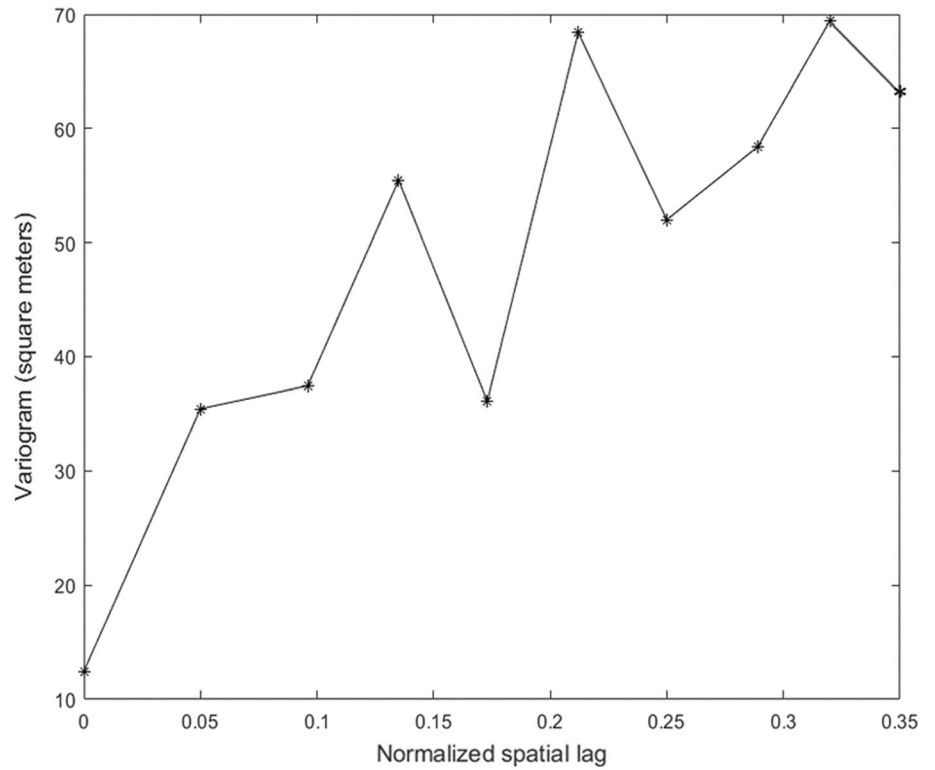
Fig. 4 Variogram fit (Spartan model) of the transformed data in terms of the entire dataset of monitoring locations in the island of Crete. The parameters of the Spartan variogram are; variogram variance: $\hat{\sigma}^2 = 44.4$ m^2, correlation length: $\hat{\xi} = 0.45$ (in normalized units), nugget: $c = 4.62$ m.^{2} and shape coefficient: $\hat{\eta}_1 = 0.56$
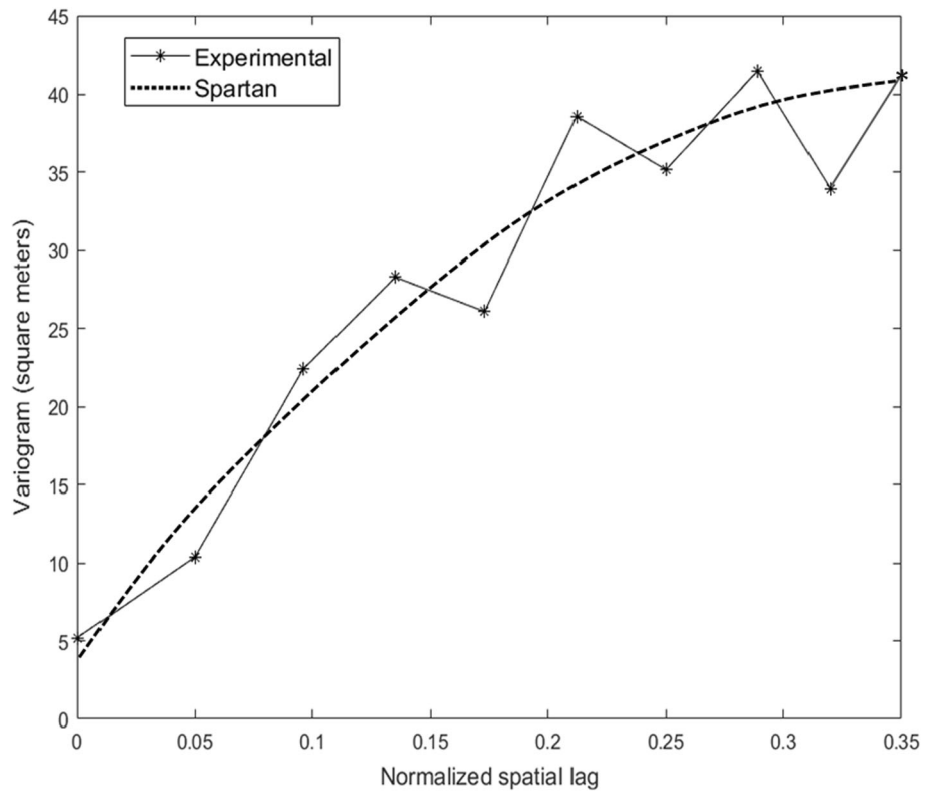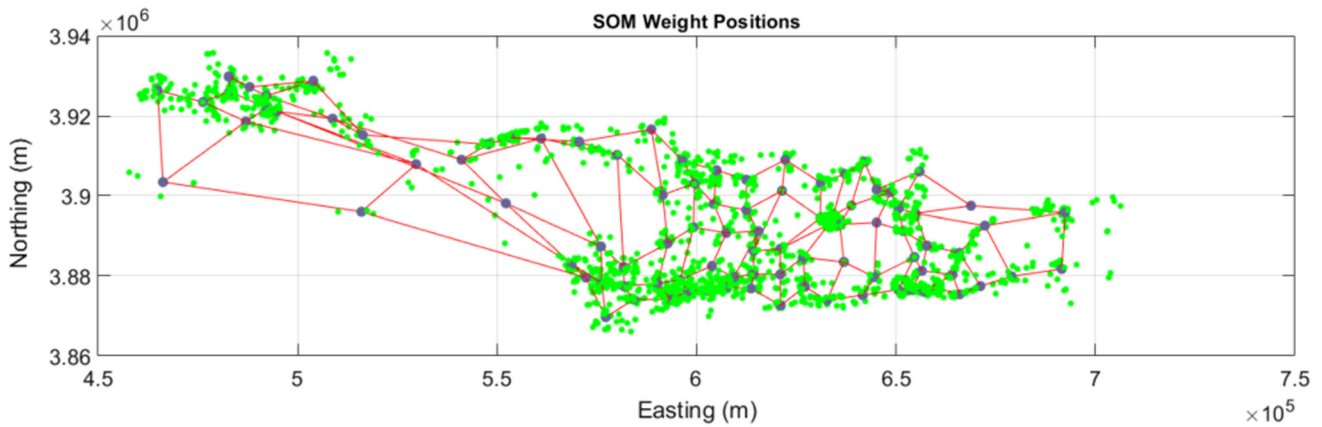
Fig. 5 Final Nodes Distribution (blue nodes) according to Self-Organizing Maps technique trained on the set of monitoring locations (light green)
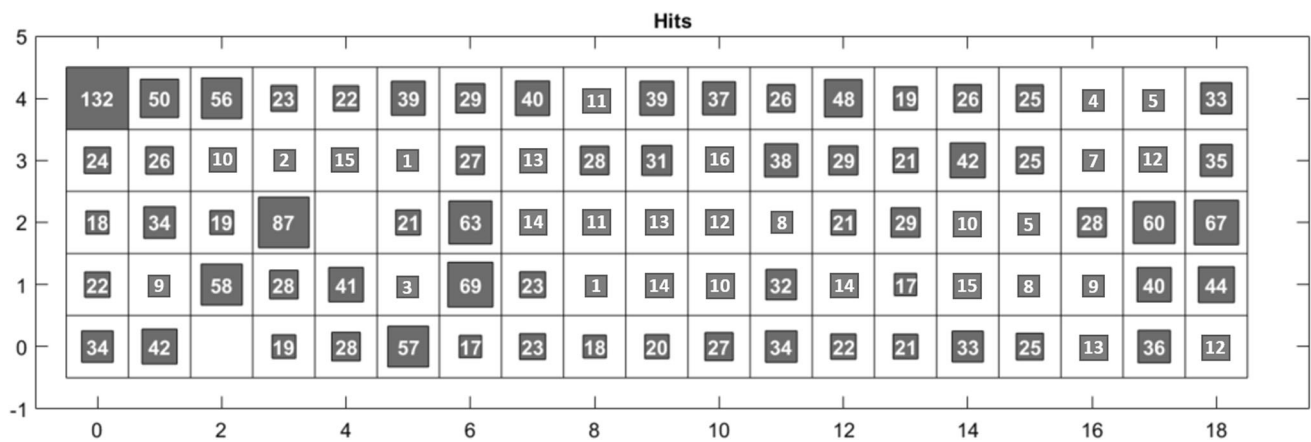
Fig. 6 Number of monitoring locations that correspond to each node

Table 2 Cross validation results of GeostatSOM-Transgaussian Kriging, Regression Kriging, Universal Kriging and Ordinary Kriging for the year 2017/18 using the Spartan variogram model and the Euclidean distance metric

Method	GeostatSOM/TGK-SOM	RK	UK	OK
Mean absolute relative error	0.11	0.20	0.24	0.34
Mean absolute error	3.09 m	5.65 m	5.93 m	7.25 m
Correlation coefficient R	0.91	0.84	0.83	0.78

calculated global variogram type (Fig. 4) using all the available measurements and its global calculated parameters are applied.

The quality of the SOM is measured by two criteria: Here, the final quantization error is 0.393, and the final topographic error is 0.013; both are very close to zero indicating optimal training.

Then, a leave-one-out cross validation analysis is performed to test the prediction efficiency of the proposed method (Table 2). The superiority of the TGK-SOM method over the classical geostatistical approaches is clear with respect to all statistical metrics, which can be explained by the more efficient selection of the estimation neighborhood with the SOM technique. The SOM clustering algorithm is an effective approach for detecting and analyzing the inherent structural modes of spatial data. It enables the interpretation of the spatial variation phenomena occurring in the study area by making the classification results for calculating the correlation neighborhood more meaningful.

Figure 7 shows the spatial distribution of the bias corrected groundwater level throughout the island of Crete using the TGK-SOM method, while Fig. 8 – presents the associated uncertainty in the estimates. On the presented
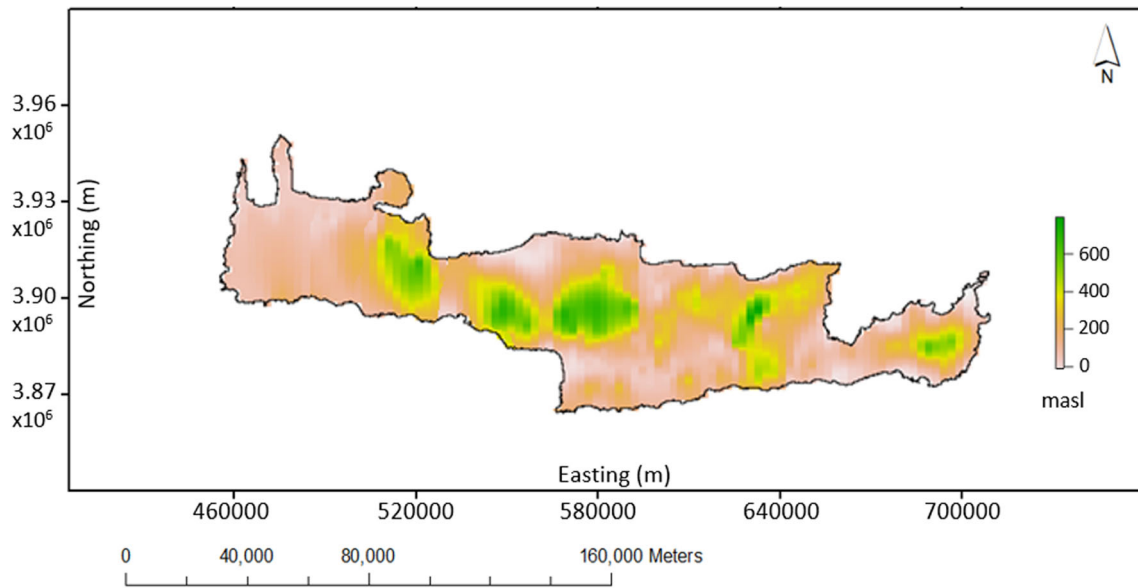
Fig. 7 Spatial distribution of bias corrected groundwater level in the island of Crete using Transgaussian Kriging by means of Self Organizing Maps
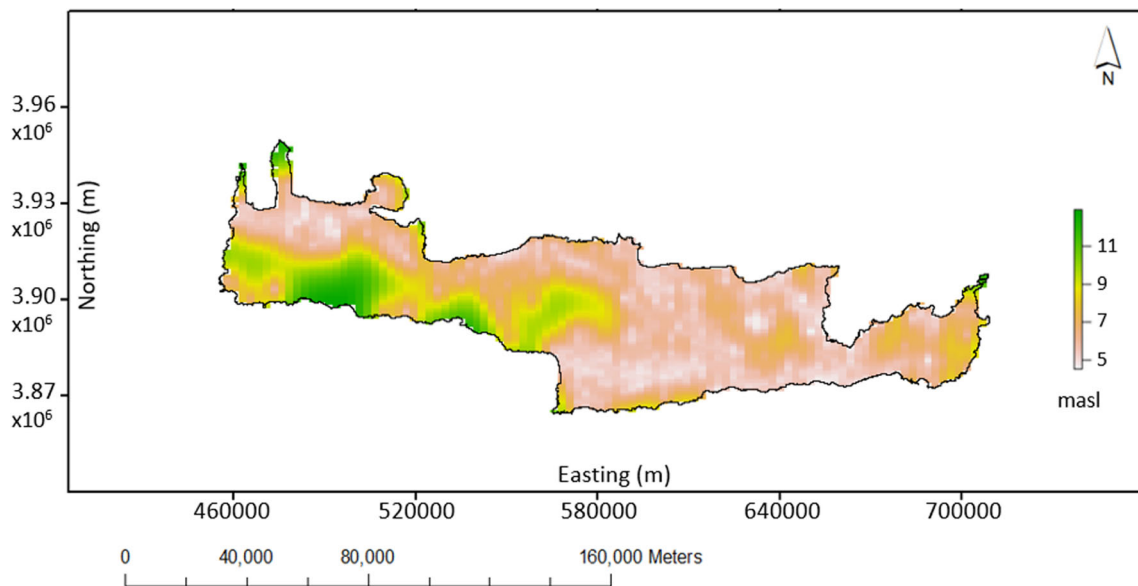

Fig. 8 Uncertainty of estimations of groundwater level in the island of Crete using Transgaussian Kriging by means of Self Organizing Maps

groundwater level spatial distribution map (Fig. 7), it can be seen that most of Crete's coastal aquifers have levels near or below sea level, indicating that salt water has intruded. Furthermore, areas with significant groundwater resources availability and those with significant shortages, primarily due to overexploitation, can be identified. Previous studies of the island's groundwater resources assessed the major aquifers separately (Special water secretariat of Greece 2017). Furthermore, the high density of data in the island's East and Northwest allows for reducing estimation uncertainty (Fig. 8) of the groundwater level distribution in inland and mountain aquifers, which can be

extremely useful for irrigation management. On the other hand, the Southwest part has fewer monitoring stations and estimates fraught with uncertainty (Fig. 8).

To further substantiate the efficiency of the proposed method, a dataset of biannual groundwater averages for the same area and year but from a different source (Decentralized Administration of Crete 2020) was used as a prediction-validation set, keeping the properties and settings of the previously applied TGK-SOM method. The new dataset consisted of 311 values spatially distributed throughout the island. The validation results were consistent with those of the leave one out cross-validation
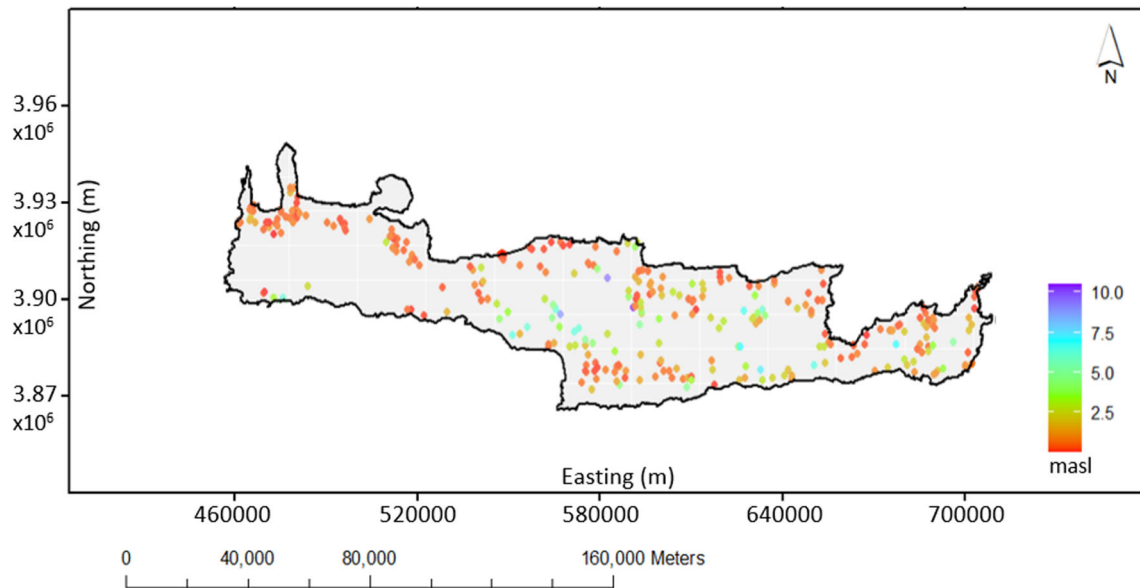
Fig. 9 Spatial distribution of mean absolute error considering the 311 monitoring locations for the hydrological year 2017/18

analysis; Mean absolute error: 3.31 m, Mean absolute relative error: 11.56%, and Correlation coefficient R: 0.90. Thus, the proposed methodology can be said to be an efficient geostatistical tool for estimating the spatial distribution of groundwater levels in complex aquifer systems.

The spatial distribution of the mean absolute error is presented in Fig. 9. Although lower the errors compared to the other methodologies, the highest errors of TGK-SOM are located mainly towards the inland of the island, in the mountainous areas, where primarily karst aquifers exist that usually deliver inconsistent measurements due to their hydrogeological nature and properties.

The cross validation results are directly comparable with other modern groundwater-level geostatistical approaches; involving Bayesian analysis (Fasbender et al. 2008; Varouchakis et al. 2019), covariates (Ruybal et al. 2019), or modelling approaches of complex hydrogeological systems (Dokou et al. 2016; Khadim et al. 2020; Tigabu et al. 2020). The cross-validation error in these works ranges from 10 to 18%. This fact validates the efficiency of the proposed methodology and its capability of exploiting efficiently information from large datasets for competent spatial analysis.

4 Conclusions

The island of Crete has a complicated network of adjacent porous and karst aquifers due to diverse boundary conditions and hydrogeological features, which prohibits a thorough physical modeling of the island's aquifers as a whole. In this study, we showed that combining geostatistics with self-organizing maps to estimate the spatial distribution of groundwater level in a large-scale application of complicated hydrogeology offers a practical and reliable new approach for this kind of applications. The groundwater level data from the entire island is used in this study, and SOM is used to take into account locally similar observations for geostatistical analysis in terms of TGK. Physically based models need a lot of data, and the proposed method offers a complementary tool.

The presented approach is to be tested on more cases, to allow for wider generalizations. Future work will be also oriented towards testing other unsupervised learning approaches of machine learning that may provide additional benefits, if compared to using self-organised feature maps, and further development of other approaches hybridizing the best features of machine learning and geostatistics.

Declarations

References

Armstrong M (1998) Basic linear geostatistics. Springer Verlag, Berlin

Augustijn E-W, Zurita-Milla R (2013) Self-organizing maps as an approach to exploring spatiotemporal diffusion patterns. Int J Health Geogr 12:60. https://doi.org/10.1186/1476-072X-12-60

Beale R, Jackson T (1990) Neural Computing-an introduction. CRC Press

Belitz K, Stackelberg PE (2021) Evaluation of six methods for correcting bias in estimates from ensemble tree machine learning regression models. Environ Modell Softw 139:105006. https://doi.org/10.1016/j.envsoft.2021.105006

Bowden GJ, Dandy GC, Maier HR (2005) Input determination for neural network models in water resources applications. Part 1—background and methodology. J Hydrol 301:75–92. https://doi.org/10.1016/j.jhydrol.2004.06.021

Box GEP, Cox DR (1964) An analysis of transformations. J R Stat Soc Ser B 26:211–252

Clark I, Harper WV (2000) Practical Geostatistics 2000. Ecosse North America Llc, Columbus, Ohio, USA

Cressie N (1990) The Origins of Kriging Math Geol 22:239–252

Cressie N (1993) Statistics for spatial data, revised. Wiley, New York

Croke B, Cleridou N, Kolovos A, Vardavas I, Papamastorakis J (2000) Water resources in the desertification-threatened Messara Valley of Crete: estimation of the annual water budget using a rainfall-runoff model. Environ Modell Softw 15:387–402

Cushman JH, Tartakovsky DM (2016) The handbook of groundwater engineering. CRC Press

Decentralized Administration of Crete (2020) Water resources portal (http://www.apdkritis.gov.gr/en/group/hydrology). Region of Crete, Directorate of Water, Heraklion

Deutsch CV, Journel AG (1992) GSLIB. Geostatistical software library and user's guide. Oxford University Press, New York

Dokou Z, Karagiorgi V, Karatzas GP, Nikolaidis NP, Kalogerakis N (2016) Large scale groundwater flow and hexavalent chromium transport modeling under current and future climatic conditions: the case of Asopos River Basin. Environ Sci Pollut Res 23:5307–5321. https://doi.org/10.1007/s11356-015-5771-1

Dowd P (2018) Quantifying the impacts of uncertainty. In: Daya Sagar BS, Cheng Q, Agterberg F (eds) Handbook of Mathematical Geosciences: Fifty Years of IAMG. Springer International Publishing, Cham, pp 349–373. https://doi.org/10.1007/978-3-319-78999-6_18

Farzad F, El-Shafie AH (2017) Performance enhancement of rainfall pattern – water level prediction model utilizing self-organizing-map clustering method. Water Resour Manag 31:945–959. https://doi.org/10.1007/s11269-016-1556-7

Fasbender D, Peeters L, Bogaert P, Dassargues A (2008) Bayesian data fusion applied to water table spatial mapping. Water Resour Res. https://doi.org/10.1029/2008WR006921

Founda D, Varotsos K, Pierros F, Giannakopoulos C (2019) Observed and projected shifts in hot extremes' season in the Eastern Mediterranean. Global Planet Change 175:190–200

García-Ruiz JM, López-Moreno JI, Vicente-Serrano SM, Lasanta Martínez T, Beguería S (2011) Mediterranean water resources in a global change scenario. Earth Sci Rev 105:121–139. https://doi.org/10.1016/j.earscirev.2011.01.006

Garrote L, Iglesias A, Granados A, Mediero L, Martin-Carrasco F (2015) Quantitative assessment of climate change vulnerability of irrigation demands in Mediterranean Europe. Water Resour Manag 29:325–338. https://doi.org/10.1007/s11269-014-0736-6

Goovaerts P (1997) Geostatistics for natural resources evaluation. Oxford University Press, New York

Hartmann A, Mudarra M, Andreo B, Marín A, Wagener T, Lange J (2014) Modeling spatiotemporal impacts of hydroclimatic extremes on groundwater recharge at a Mediterranean karst aquifer. Water Resour Res 50:6507–6521. https://doi.org/10.1002/2014WR015685

Henriques R, Bacao F, Lobo V (2012) Exploratory geospatial data analysis using the GeoSOM suite. Comput Environ Urban Syst 36:218–232. https://doi.org/10.1016/j.compenvurbsys.2011.11.003

Hoogland T, Heuvelink GBM, Knotters M (2010) Mapping water-table depths over time to assess desiccation of groundwater-dependent ecosystems in the Netherlands. Wetlands 30:137–147

Hsu K-C, Li S-T (2010) Clustering spatial–temporal precipitation data using wavelet transform and self-organizing map neural network. Adv Water Resour 33:190–200. https://doi.org/10.1016/j.advwatres.2009.11.005

Huang Y, Ye H, Zhang L, Zhang S, Shen C, Li Z, Huang Y (2017) Prediction of soil organic matter using ordinary kriging combined with the clustering of self-organizing map: a case study in Pinggu District, Beijing. China. Soil Sci 182(2):52–62

Kalteh AM, Hjorth P, Berndtsson R (2008) Review of the self-organizing map (SOM) approach in water resources: analysis, modelling and application. Environ Modell Softw 23:835–845. https://doi.org/10.1016/j.envsoft.2007.10.001

Kanevski M (2013) Advanced mapping of environmental data. John Wiley & Sons

Khadim FK, Dokou Z, Lazin R, Moges S, Bagtzoglou AC, Anagnostou E (2020) Groundwater modeling in data scarce aquifers: the case of Gilgel-Abay, Upper Blue Nile, Ethiopia. J Hydrol 590:125214. https://doi.org/10.1016/j.jhydrol.2020.125214

Kim G-U, Seo K-H, Chen D (2019) Climate change over the Mediterranean and current destruction of marine ecosystem. Sci Rep 9:18813. https://doi.org/10.1038/s41598-019-55303-7

Kitanidis PK (1997) Introduction to geostatistics. Cambridge University Press, Cambridge

Kohonen T (2001a) Applications. In: Kohonen T (ed) Self-Organizing Maps. Springer, Berlin, pp 263–310. https://doi.org/10.1007/978-3-642-56927-2_7

Kohonen T (2001b) The basic SOM. In: Kohonen T (ed) Self-Organizing Maps. Springer, Berlin, pp 105–176

Kohonen T (2013) Essentials of the self-organizing map. Neur Net 37:52–65. https://doi.org/10.1016/j.neunet.2012.09.018

Koutroulis AG, Tsanis IK, Daliakopoulos IN, Jacob D (2013) Impact of climate change on water resources status: A case study for Crete Island, Greece. J Hydrol 479:146–158. https://doi.org/10.1016/j.jhydrol.2012.11.055

Krige DG (1951) A statistical approach to some basic mine valuation problems on the Witwatersrand. J Chem Metall Min Soc South Africa 52:119–139

Manzione RL, Castrignanò A (2019) A geostatistical approach for multi-source data fusion to predict water table depth. Sci Total Environ 696:133763. https://doi.org/10.1016/j.scitotenv.2019.133763

Markonis Y et al (2018) Global estimation of long-term persistence in annual river runoff. Adv Water Resour 113:1–12. https://doi.org/10.1016/j.advwatres.2018.01.003

Matheron G (1971) The theory of regionalized variables and its applications. Ecole Nationale Superieure des Mines de Paris, Fontainebleau, Paris

Matiatos I, Wassenaar LI (2019) Stable isotope patterns reveal widespread rainy-period-biased recharge in phreatic aquifers across Greece. J Hydrol 568:1081–1092. https://doi.org/10.1016/j.jhydrol.2018.11.053

Nourani V, Baghanam AH, Adamowski J, Gebremichael M (2013) Using self-organizing maps and wavelet transforms for space–time pre-processing of satellite precipitation and runoff data in neural network based rainfall–runoff modeling. J Hydrol 476:228–243. https://doi.org/10.1016/j.jhydrol.2012.10.054

Panagiotou CF, Kyriakidis P, Tziritis E (2022) Application of geostatistical methods to groundwater salinization problems: a review. J Hydrol 615:128566. https://doi.org/10.1016/j.jhydrol.2022.128566

Ribeiro PJ Jr, Diggle PJ, Ribeiro MPJ Jr, Suggests M (2007) The geoR package. R News 1:14–18

Richardson AJ, Risien C, Shillington FA (2003) Using self-organizing maps to identify patterns in satellite imagery. Prog Oceanogr 59:223–239. https://doi.org/10.1016/j.pocean.2003.07.006

Ruybal CJ, Hogue TS, McCray JE (2019) Evaluation of groundwater levels in the arapahoe aquifer using spatiotemporal regression kriging. Water Resour Res. https://doi.org/10.1029/2018wr023437

Schabenberger O, Gotway CA (2005) Statistical Methods for Spatial Data Analysis. CRC Press, Boca Raton

Seifeddine J et al. (2021) Multidisciplinary joint-force efforts towards science-based management in the Mediterranean region a particular focus on transboundary aquifers Paper presented at the Transboundary Aquifers Challenges and the way forward, ISARM2021, UNESCO, Paris

Smerdon BD (2017) A synopsis of climate change effects on groundwater recharge. J Hydrol 555:125–128. https://doi.org/10.1016/j.jhydrol.2017.09.047

Special water secretariat of Greece (2017) Integrated Management Plans of the Greek Watersheds, River basin management report for the water sector of Crete (in Greek). Ministry of Environment & Energy, Athens

Special water secretariat of Greece (2020) National Water Monitoring Network, groundwater data (in Greek) - http://nmwn.ypeka.gr/?q=groundwater-stations. Athens, Greece, Access Date: 20/10/2020

Theodoridou PG, Varouchakis EA, Karatzas GP (2017) Spatial analysis of groundwater levels using Fuzzy Logic and geostatistical tools. J Hydrol 555:242–252. https://doi.org/10.1016/j.jhydrol.2017.10.027

Thomas BF, Famiglietti JS (2019) Identifying climate-induced groundwater depletion in GRACE observations. Sci Rep 9:4124. https://doi.org/10.1038/s41598-019-40155-y

Tigabu TB, Wagner PD, Hörmann G, Fohrer N (2020) Modeling the spatio-temporal flow dynamics of groundwater-surface water interactions of the Lake Tana Basin, Upper Blue Nile, Ethiopia. Hydrol Res 51:1537–1559. https://doi.org/10.2166/nh.2020.046

Toth E (2009) Classification of hydro-meteorological conditions and multiple artificial neural networks for streamflow forecasting. Hydrol Earth Syst Sci 13:1555–1566. https://doi.org/10.5194/hess-13-1555-2009

Varouchakis EA (2016) Integrated water resources analysis at basin scale: a case study in Greece. J Irrig Drain E-ASCE 142:05015012. https://doi.org/10.1061/(ASCE)IR.1943-4774.0000966

Varouchakis EA (2021) Gaussian transformation methods for spatial data. Geosciences 11:196

Varouchakis EA, Hristopulos DT (2013) Improvement of groundwater level prediction in sparsely gauged basins using physical laws and local geographic features as auxiliary variables. Adv Water Resour 52:34–49

Varouchakis EA, Hristopulos DT (2019) Comparison of spatiotemporal variogram functions based on a sparse dataset of groundwater level variations. Spat Stat 34:100245. https://doi.org/10.1016/j.spasta.2017.07.003

Varouchakis EA, Hristopulos DT, Karatzas GP (2012) Improving kriging of groundwater level data using nonlinear normalizing transformations-a field application. Hydrolog Sci J 57:1404–1419

Varouchakis EA, Corzo GA, Karatzas GP, Kotsopoulou A (2018) Spatio-temporal analysis of annual rainfall in Crete, Greece. Acta Geophys 66:319–328. https://doi.org/10.1007/s11600-018-0128-z

Varouchakis EA, Theodoridou PG, Karatzas GP (2019) Spatiotemporal geostatistical modeling of groundwater levels under a Bayesian framework using means of physical background. J Hydrol 575:487–498. https://doi.org/10.1016/j.jhydrol.2019.05.055

Wada Y, van Beek LPH, van Kempen CM, Reckman JWTM, Vasak S, Bierkens MFP (2010) Global depletion of groundwater resources. Geophys Res Lett. https://doi.org/10.1029/2010GL044571

Wehrens R, Buydens LM (2007) Self-and super-organizing maps in R: the Kohonen package. J Stat Softw 21:1–19