Hydraulic head uncertainty estimations of a complex artificial intelligence model using multiple methodologies

E. Tapoglou, E. A. Varouchakis, I. C. Trichakis and G. P. Karatzas

ABSTRACT

The purpose of this study is to examine the uncertainty of various aspects of a combined artificial neural network (ANN), kriging and fuzzy logic methodology, which can be used for the spatial and temporal simulation of hydraulic head in an aquifer. This simulation algorithm was applied in a study area in Miami – Dade County, USA. The percentile methodology was applied as a first approach in order to define the ANN uncertainty, resulting in wide prediction intervals (PIs) due to the coarse nature of the methodology. As a second approach, the uncertainty of the ANN training is tested through a Monte Carlo procedure. The model was executed 300 times using different training set and initial random values, and the training results constituted a sensitivity analysis of the ANN training to the kriging part of the algorithm. The training and testing error intervals for the ANNs and the kriging PIs calculated through this procedure can be considered narrow, taking into consideration the complexity of the study area. For the third and final approach used in this work, the uncertainty of kriging parameter was calculated through the Bayesian kriging methodology. The derived results prove that the simulation algorithm provides consistent and accurate results.

Key words | artificial neural networks, Bayesian uncertainty, fuzzy logic, kriging, uncertainty analysis

E. Tapoglou (corresponding author)
E. A. Varouchakis
I. C. Trichakis
G. P. Karatzas
School of Environmental Engineering,
Technical University of Crete,
Chania 73100,
Greece
E-mail: *etapoglou@isc.tuc.gr*

INTRODUCTION

Data-driven models and statistical models have both been extensively used in groundwater modeling, especially in the last decades (Delbari *et al.* 2014; Maiti & Tiwari 2014). A new concept, presented in Tapoglou *et al.* (2014), allowed for the combination of two separate methodologies, artificial neural networks (ANNs) and kriging for the spatial and temporal simulation of hydraulic head using a fuzzy logic system in a 7,800 m^2 area covering Bavaria region, Germany. In this study, to examine the uncertainty of various aspects of the combined methodology and to prove the effectiveness of this approach providing better simulation results, the same methodology was applied in a smaller study area in Miami, Dade County, USA (1,460 m^2). Preliminary results of this study are presented in Tapoglou *et al.* (2018).

To provide a measure of the uncertainty, an analysis comprising of three different methodologies was performed. The first methodology corresponded to the uncertainty of kriging parameters, while the remaining two estimated the uncertainty derived from the ANNs. The remainder of this paper is organized as follows: the last part of the introduction presents briefly the simulation algorithm while in the next section the methodology is described, followed by some general information about the study area and the numerical results of the methodology. The last part includes the most important findings and the conclusion of this study.

Simulation algorithm

The simulation algorithm used in this work is an evolution of the one presented in Tapoglou *et al.* (2014) and involves the use of ANNs, kriging and a fuzzy logic methodology in order to simulate the water table changes in a temporal and spatial scale. The main idea of the conceptual

doi: 10.2166/hydro.2019.137

framework is that the ANN simulates the temporal changes in the hydraulic head, while kriging interpolates the results in a spatial manner to cover the whole area, creating a two-dimensional map from a set of point measurements.

The temporal simulation took place in multiple locations where data were available, diversifying the inputs as necessary in order to acquire the optimal ANN results. Owing to these different needs for input parameters and in order to get the best possible results, one ANN was developed for every location with available hydraulic head data. Many studies have followed a similar approach in the past, since it provides a higher flexibility in input parameter selection, combined with reduced computational effort for training compared to creating a single ANN for all the outputs (Sahoo *et al.* 2017).

Data-driven methods, like ANNs in this study, require the existence of a large dataset of input and target values. These prerequisites include information on meteorological and hydrological parameters measured on a daily basis (same as the time step in this study) as input data and hydraulic head measurements in the same time step as the target values of the output parameter during training. The selection of input parameters is focused on those who are linked directly or indirectly to the water budget. The ANNs used available data both for training and simulation of the current state, while the trained ANNs are able to predict future status of the aquifer, when evaluating different scenarios.

The results of all ANNs were subsequently interpolated using the kriging methodology. Ordinary kriging had been used, and the linear, exponential and power-law variograms were tested. The interpolation was conducted in a grid inside the convex polygon of the wells. This grid is a rectilinear grid, where all observation points fall exactly at a node. The nodes of the grid where observed data are not available are the prediction points. To exclude from the simulation grid nodes that fall far outside the study area and far from any observation point, Delaunay triangles were created from the known data locations. Nodes that fell outside the area covered by at least one triangle were not considered as prediction points.

To improve the initial results, a fuzzy logic system was used as a means to combine the two methodologies. More specifically, the neighborhood used in kriging interpolation was defined through fuzzy logic, combining the distance between the prediction point and data points with the ANN training and testing error in every location. For every prediction point involved in the training and testing process and for every time step, a different variogram is calculated and the hydraulic head is simulated. A detailed description of the functionality of the original methodology, followed also in this study, is found in Tapoglou *et al.* (2014). The main enhancement of the methodology in this study consists of the improved selection of ANN input parameters to reduce ANN error and the subsequently applied metamodel for uncertainty quantification of the different components of this study's methodology.

METHODOLOGIES

The methodologies followed in the present paper will cover the uncertainties involved in the simulation algorithm described in Tapoglou *et al.* (2014). Model uncertainty can be attributed to many reasons. In this work, we have identified and grouped the sources of uncertainty in two main categories, uncertainty derived from the ANN training process and uncertainty due to the kriging interpolation methodology.

In the past, various methodologies have been used in different applications for the calculation of model uncertainty. Monte Carlo simulation and bootstrap methods are very common and can give a perspective of the uncertainty involved. The objective of an uncertainty analysis is to show the effects of model inputs on the model simulation results (Eslamian 2014). In this study, different methodologies were used for the determination of uncertainty derived from different parts of the developed algorithm (Figure 1). As far as the ANN part of the simulation algorithm is concerned, two methodologies were followed: the percentile methodology and a Monte Carlo simulation for the uncertainty of the ANN's training. The percentile methodology is simple and easy to apply during ANN training, but studies have shown that it produces larger confidence intervals (Jiang *et al.* 2013). The second methodology produces narrower intervals which take into account the value and individual characteristics of each data point, producing more meaningful results. These results later became the

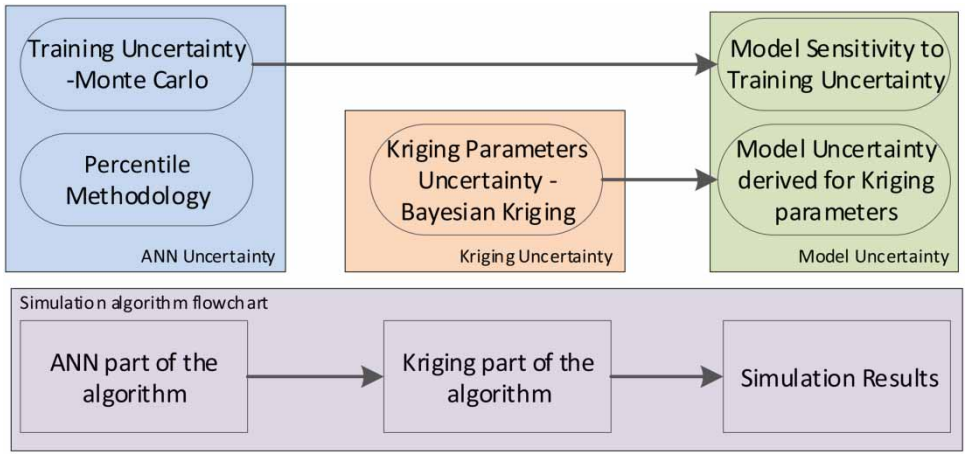

Figure 1 | Methods tested in each simulation step.

input for the Kriging part of the algorithm, constituting a sensitivity analysis of the kriging interpolation method to the ANN training. ANN training uncertainty can be attributed to a number of factors ranging from initialization of the weights and other processes involving random number generators. Uncertainty due to the Kriging parameters is also studied using the Bayesian kriging methodology.

Uncertainty in ANN results

Uncertainty due to the ANN model has been extensively studied in the past, using both bootstrap-based methods (Trichakis *et al.* 2011; Kasiviswanathan & Sudheer 2013) and Monte Carlo simulation methods (Jiang *et al.* 2013; Dehghani *et al.* 2014).

In ANNs, the source of uncertainty may be attributed to two reasons, model structure and training dataset (Dybowski & Roberts 2001). In this study, the model structure is fixed and only the uncertainty derived from the training dataset is considered. This uncertainty can be attributed to not only the imperfection in data collection tools or techniques, but also to the available amount of training data and how representative of the reality they are. Moreover, not all the possible realizations of the variables involved are available in the training set, as well as the training set itself is only one of a large number of possibilities. For example, the training dataset can be short or might not contain extreme values that are nevertheless possible to occur in the area. This will increase the uncertainty of the model when it is extrapolating.

By using these two methodologies, the prediction intervals (PIs) can be calculated.

Percentile prediction intervals

In this approach, the PIs can be calculated through the statistical characteristics of the model errors that occurred when reproducing those observed in the field data (Shrestha & Solomatine 2006).

This approach takes a large number of simulated variables of interest and orders them from smallest to largest. A 90% PI is then computed by identifying the lower PI as the 5th percentile and the upper PI as the 95th percentile. This leaves 90% of the simulated estimates within this range, while dividing the remaining 10% of the simulated values equally into the upper and lower tails. The advantage of this method is that it does not make any distributional assumptions and it does not require the distribution to be symmetric. However, evaluating the PI using this methodology yields larger confidence intervals compared to other methodologies.

In order to use this methodology to evaluate every ANN's PIs, for every ANN involved separately, the difference between the real and the simulated values is calculated for every time step studied. These differences are then sorted, and the 5th and 95th percentile values are identified. The values are then added to all simulated outputs, concluding in two new time series which represent the 5% upper and lower limit of the prediction.

Monte Carlo method

Conventionally, a single ANN is developed through training, and once completed, the ANN parameters are considered fixed. This way, an ANN becomes a deterministic model; a given input data vector will always produce a specific output value. To calculate the uncertainty in this prediction, Monte Carlo methods can be used, which are typically computer-based techniques for brute force numerical simulation of probabilistic processes.

The first step in this methodology is to generate a large number of sets of random numbers that have statistical properties similar to those of the real variables. If there are actual measurements describing the variables, it is also possible to select randomly among the measured values, as an approximation to the distribution underlying the particular values measured in the field. Using each set of values as input variables, the output variable is calculated. This process is then repeated several times to test whether the conclusions drawn are sensitive to the particular random values that have been generated. The number of repetitions necessary to achieve a particular level of precision is problem-dependent. The more iterations are conducted, the more accurate the prediction intervals will be. However, the computational costs may be enormous for a large number of iterations.

The uncertainty of the ANN results in this case can be assessed mainly by the random initialization of values (i.e. the ANN synaptic weights), as well as in the random selection of training and testing datasets. To investigate the uncertainty in the predictions of the ANN, multiple different executions of the algorithm, with random initial values for the neural weights and different training and testing sets, were performed. All the random number generators were set to choose a random integer number with uniform distribution. The examination of the uncertainty attributed to the architecture and structure of the ANNs would require large computational cost; hence, it is not examined in this study.

Uncertainty in kriging results

The uncertainty of the kriging algorithm is evaluated through the Bayesian kriging methodology, which can be used to estimate the uncertainty attributed to the kriging parameters.

Bayesian kriging

The process of interpolating spatial data using kriging can be realized in two steps. First, the covariance function/semivariogram is calculated and then the kriging interpolation takes place. When using the classical approach, the assumption that the model parameters are known is made, and the uncertainty in the calculation of variogram parameters (sill and range) is considered as minimal. The purpose of the Bayesian kriging is to quantify the uncertainty of these parameters and hence the uncertainty of the covariance estimation.

Typically, using Bayes theory, one can acquire a posteriori distribution of parameters using the prior distribution and the initial values of the parameters. Using a posteriori distribution, a prediction of the parameters can be done. In this way, a model 'averaging' over the unknown model parameters is constructed.

However, when using spatial data, the prior cannot be calculated easily. In Bayesian kriging (Pilz & Spöck 2008), the posterior distribution of the parameters involved (semivariogram and/or transformation parameters) is specified by means of simulations, which correspond to the uncertainty of covariance parameters.

To implement Bayesian kriging methods, first the appropriate variogram model must be fitted to the experimental data, using the standard procedure followed when using the variogram estimation (first step in kriging interpolation methodology). Using the parameters derived and the covariance function of the model, the covariance matrix is constructed.

A common method for testing a statistical model is the use of artificial data. In order to do so, the statistical properties have to be embedded in the dataset and then the model must be examined for the presence of these effects and how they behave under different experimental conditions.

Uniform random numbers generation is the first step in this procedure. A generally approved method of correlating random numbers with a known covariance matrix (C) is by finding a matrix U, according to the following equation:

$$U^T \cdot U = C \tag{1}$$

Matrix U is a triangular matrix and can be derived using decomposition methods. Matrix decomposition is commonly used in the Monte Carlo method for simulating systems with multiple correlated variables.

Using this matrix, correlated random numbers R_c can be generated from uncorrelated numbers R, by multiplying them with this matrix in the following equation:

$$R_c = R \cdot U \tag{2}$$

These correlated random values have the same statistical characteristics with the initial, observed in the field, data. These random values are then fitted into a variogram and the value in an unknown location is estimated using kriging.

Cholesky decomposition is one way of providing these random correlated values. The covariance matrix of a vector Y can be given as follows: $C = E(YY^T)$. If X is a random vector, consisting of uncorrelated random values uniform in [0,1], then $E(XX^T) = I$.

The Cholesky decomposition of the correlation matrix is given as $C = L \cdot L^T$. Note that it is possible to obtain a Cholesky decomposition of C since by definition the covariance matrix C is symmetric and positive definite.

When the random vector X is multiplied by L ($Z = L \cdot X$), statement (3) can be formed:

$$\begin{aligned} E(Z \cdot Z^T) &= E((L \cdot X) \cdot (L \cdot X^T)) = E(L \cdot X \cdot X^T \cdot L^T) \\ &= L \cdot E(X \cdot X^T) \cdot L^T = L \cdot I \cdot L^T = L \cdot L^T = C \end{aligned} \tag{3}$$

In this way, it can be proved that the random vector Z has the desired covariance matrix C.

The distribution of the simulated values using the kriging of different correlated random values can be used in order to derive the confidence intervals of the process. In this way, the distribution of the parameters involved (sill, range etc.) in the process is used as a prior knowledge, while the posterior distribution reflects the uncertainty of the covariance estimation (Pilz & Spöck 2008).

Uncertainty in simulation algorithm results

The overall simulation algorithm uncertainty was evaluated following two methodologies. The first one included the use of the Monte Carlo results of the ANN uncertainty analysis and can also be characterized as a sensitivity analysis of the kriging to the ANN training. The results of the analysis described in the section 'Bayesian kriging' can also serve as uncertainty analysis of the simulation algorithm to the kriging parameter estimation process (see Figure 1).

Monte Carlo method

In this case, the results of the Monte Carlo simulation in training the ANNs were used. More specifically, the multiple different training results (different neural weights) were used for the evaluation of the hydraulic head, resulting in multiple different values for every time step and for every prediction point. Using the 5th and 95th percentile, it is possible to determine the 90% prediction interval. These results were then passed forward to the kriging part of the algorithm, generating the range the simulation results are within, for every prediction point and time step separately.

STUDY AREA

The area in which the proposed methodology is applied is located in Miami, Dade County, Florida, USA. A total number of 30 wells within Biscayne aquifer inside or around the urban area of Miami were used for the water table simulation. The available data covered the time period from 25 April 2010 to 22 February 2014, with daily time step, and included the hydraulic head, the discharge and water level in four nearby rivers and meteorological data from two stations. The data used in this study regarding water levels were acquired through United States Geological Survey (USGS), while the meteorological data from the National Climatic Data center of the National Oceanic and Atmospheric Administration (NOAA). All the data used were directly or indirectly linked to the aquatic equilibrium. The locations of (a) 30 hydraulic head measuring stations, (b) four surface water stations and (c) two meteorological stations are depicted in Figure 2.

Biscayne aquifer is one of the most productive aquifers in the world; it is unconfined and shallow with the average hydraulic head often being less than 1 m. It is wedge-shaped, extending from a few meters thickness at the west coast to as much as 45 m thick near the east coast. As analyzed by USGS (2014), geologically, the study area is dominated by Miami Limestone and Key Largo Limestone. It consists of

Figure 2 | (a) Hydraulic head measuring stations, (b) surface water measuring stations and (c) meteorological stations.

very porous limestone, created by secondary dissolution (Fish & Stewart 1991).

Water supply and water management systems, like regulated channels, can affect the spatial correlation of water level between monitoring wells. Natural factors can affect the spatial correlation of the data for the groundwater level. One such natural factor, rainfall, can be intense and extremely localized. Therefore, distant wells are less likely to show influence from the same rainfall events. Land use is another factor that could affect correlation. Rainfall-runoff in urbanized areas is different than in natural settings. In urban settings, pavement and storm sewers direct water into streams and channels; in natural settings, rainfall can readily infiltrate into the soil. Evapotranspiration in urban settings is also likely to be different than the one in a rural area. Because of the combined effect of all of these factors, correlation of the water level data in distant wells is generally expected to be small.

RESULTS

Preprocessing of available data

An important factor that can easily be neglected when choosing ANN parameters is that of the time lag, i.e. the time between an event happening (for example, a rainfall occurence) and the appearance of its effect in the hydraulic head dataset. When rainfall and change of groundwater level are the parameters under consideration, this lag represents the time that it takes for the water to penetrate the soil and reach the aquifer. The correlation coefficient between time series A for a parameter and time series B for the hydraulic head change can be used to determine the optimal lag between the parameter and the hydaulic head change in the following equation:

$$\text{Correl}(A,B) = \frac{\sum (a-\bar{a})\cdot(b-\bar{b})}{\sqrt{\sum (a-\bar{a})^2 \cdot (b-\bar{b})^2}} \quad (4)$$

where a denotes the value of the parameter, b denotes the value of the hydraulic head change and $\bar{a}$, $\bar{b}$ are the average values of time series A and B, respectively.

The selection of the lag of each parameter was determined by the highest correlation between input and output of the ANN. Parameters considered to have time lag include precipitation (rainfall, snowfall, etc.) and parameters related to surface water recharge or discharge (water level in a lake or river, flow rate, etc.). More than one time lag can have approximately equal correlation coefficient between an input parameter and the hydraulic head change.

The selection of input parameters relies on the outcome of the correlation coefficient analysis. In some cases, the time lag between the input and the output parameter is

apparent, and the input parameter with the appropriate lag (the one which has the highest correlation) can enter the model. In other cases, the effect of the input parameter is apparent in more than 1 day. This is particularly true for rainfall, since the effect of a rainfall event, due to the relatively small groundwater velocities, is visible for more than 1 day, creating a hydrograph, similar to the one describing surface runoff. This effect is reflected in the values of the correlation coefficients between the input and the output variables (in this case, rainfall and hydraulic head change). These coefficients get their highest values in a range of consecutive days rather than in a single day with a single time lag. Therefore, for these parameters, all the values of the input parameter with a different time lag that had an effect on the output constituted part of the input dataset. The input parameters to the ANNs were a combination of the parameters with the highest correlation coefficient. The ANN output parameter is always the hydraulic head.

For each one of the ANNs, the following input parameters were used:

- Precipitation for two meteorological stations (3 days)
- Soil temperature
- Air temperature
- Humidity
- Surface water levels from two stations (2 days)
- Discharge from one station (best day)

In most occasions, the best time lag, in terms of correlation coefficient, was between 0 and 1 day for the surface water level and discharge, and for the rainfall, this time lag varied from 0 to 3 days, with the most common being 0 and 1 day. These results are to be expected in this study area, since the aquifer is highly productive, with water flowing through easily. Therefore, a small time lag of 0–1 days and at most 3 days also seems to have a physical meaning. The correlation coefficient values, of the time lagged parameters in respect to the output parameter that were finally used, varied from 0.4 to 0.8.

There is not a universal rule about the optimal architecture of an ANN. However, according to the rule of Fine (1999), using three times as many training patterns as network parameters (weights) is adequate to achieve good generalization. Taking into account that nine input parameters, one output and 1,100 training sets were used, two architectures were examined; the first one with one hidden layer with 32 hidden nodes (Architecture 1) and the second one with two hidden layers with 17 and 12 hidden nodes, respectively (Architecture 2).

Three variogram models were also tested for the best architecture, the linear, the exponential and the power-law, in order to choose the most appropriate for the data at hand. These variograms were chosen in order to have one variogram model per type: the simplest models, the linear, one differentiable, the exponential, and one non-differentiable, the power-law.

The root-mean-squared error (RMSE) and the Nash-Sutcliffe efficiency (NSE) coefficient (Equation (5)) were used as a measure of error.

$$\mathrm{NSE} = 1 - \frac{\sum_{t=1}^{T} (Q_{\mathrm{m}}^{t} - Q_{\mathrm{o}}^{t})^2}{\sum_{t=1}^{T} (Q_{\mathrm{o}}^{t} - \overline{Q_{\mathrm{o}}})^2} \tag{5}$$

where Q_{o}^{t} is the observed value at time step t, and Q_{m}^{t} is the simulated value at the same time step.

Simulation results

The ANN performances for each one of the two possible architectures (Architecture 1 – one hidden layer with 32 hidden nodes and Architecture 2 – two hidden layers with 17 and 12 hidden nodes) are presented in terms of RMSE in Table 1.

In terms of average minimum and maximum values, the second architecture has better results, both for the training and testing error. To evaluate the performance of each one of the 30 ANNs developed, they were divided into three categories, depending on their training and testing RMSE performance (low, medium and high RMSE values). The

Table 1 | Training and testing RMSE for two architectures

RMSE (m)	Architecture 1	Architecture 2
Average training error	1.16×10^{-3}	6.63×10^{-4}
Average testing error	1.20×10^{-3}	7.92×10^{-4}
Maximum training error	2.19×10^{-3}	1.78×10^{-3}
Maximum testing error	3.42×10^{-3}	2.98×10^{-3}
Minimum training error	2.46×10^{-4}	2.23×10^{-4}
Minimum testing error	2.63×10^{-4}	1.26×10^{-4}

results will be presented for one representative well for each category. The training RMSE, testing RMSE and the overall NSE coefficient for the representative well of each category are summarized in Table 2.

The first category includes the ANNs with the best performance, represented by well 11. The second category has average results, represented by well 3, while the third category has the worst RMSE results amongst the ANNs studied. It should be noted that the NSE values are not very close to the optimal value, which equals one ($NSE_{optimal} = 1$). This can be attributed to the low hydraulic head values that should be simulated by the model. As indicated by Schaefli & Gupta (2007), it is not possible to achieve very high NSE values when simulating small absolute values; however, it is an indication of the comparative performance of the model. Nevertheless, the NSE values are well above zero, which means that our model describes much better the output than a simple average output value could have.

The simulation results as hydraulic head for the three representative wells are depicted in Figure 3.

Next, the algorithm, with the use of fuzzy logic, defines the appropriate neighbors, which will be used for the kriging interpolation. The number of neighbors for each prediction point is set to 20. This number should be chosen carefully so as to allow for the 30 pairs per distance class, which, according to Journel & Huijbregts (1978), are necessary in order to acquire a good fitting of the experimental variogram to the theoretical model.

To evaluate the results, cross-validation was performed for all three variograms studied. For every data point, for 10% of the time steps, the observed values at that data point were ignored, and the algorithm follows as if the data point does not exist. In this way, three values are available: the observed in the field, the ANN simulated and the kriging simulated. In the end of the process, the kriging simulated values are compared to the real and ANN simulated, in order to derive an independent measure of performance. In terms of three error indicators, RMSE, mean absolute error (MAE) and bias, the results are 1.14, 0.89 and −0.0485 for the linear variogram, 0.962, 0.7047 and 0.0499 for the exponential and 0.98, 0.7422 and 0.0446 for the power correspondingly.

For all three error indicators used, the exponential variogram yielded better results and will be used henceforth. The cross-validation results can be summarized in Figure 4, where observed values are plotted against the simulated hydraulic head. The closer the simulation results are in the $x = y$ line, the better the simulation is.

As it is apparent in Figure 4, the kriging interpolation tends to correct any misjudgments of the hydraulic head performed by the ANNs. The improvement in the ANN results is also reflected in the error indicators. More specifically, the RMSE, NSE and MAE values of the data used for the cross-validation when only using ANNs are 0.58, 0.59 and 0.77, while when using ANNs together with kriging are 0.8, 0.41 and 0.49, respectively.

Uncertainty results

ANN prediction confidence intervals

Using the initial simulation results, it is possible to calculate the maximum difference between observed and simulated values. By applying the 95% and 5% percentile error to all simulated values, the 95% and 5% percentile of the simulation results can be calculated. The prediction interval width for the three representative wells is 1.18 (−0.43 to 0.75), 1.15 (−0.59 to 0.56) and 1.29 (−0.58 to 0.71) for wells 11, 3 and 18 correspondingly. A noteworthy fact is that with this methodology, each well gets a different width of prediction interval, and the fact that while the results for one well may be better in terms of RMSE, there is no guarantee that the interval width will also be narrower. This width

Table 2 | Performance of the three representative wells

Category number	Representative well	Training error of the representative well (m)	Testing error of the representative well (m)	NSE of the representative well	Number of wells within the category
1	11	2.77×10^{-4}	3.13×10^{-4}	0.69	9
2	3	5.68×10^{-4}	4.19×10^{-4}	0.65	13
3	18	1.11×10^{-3}	6.23×10^{-4}	0.58	8

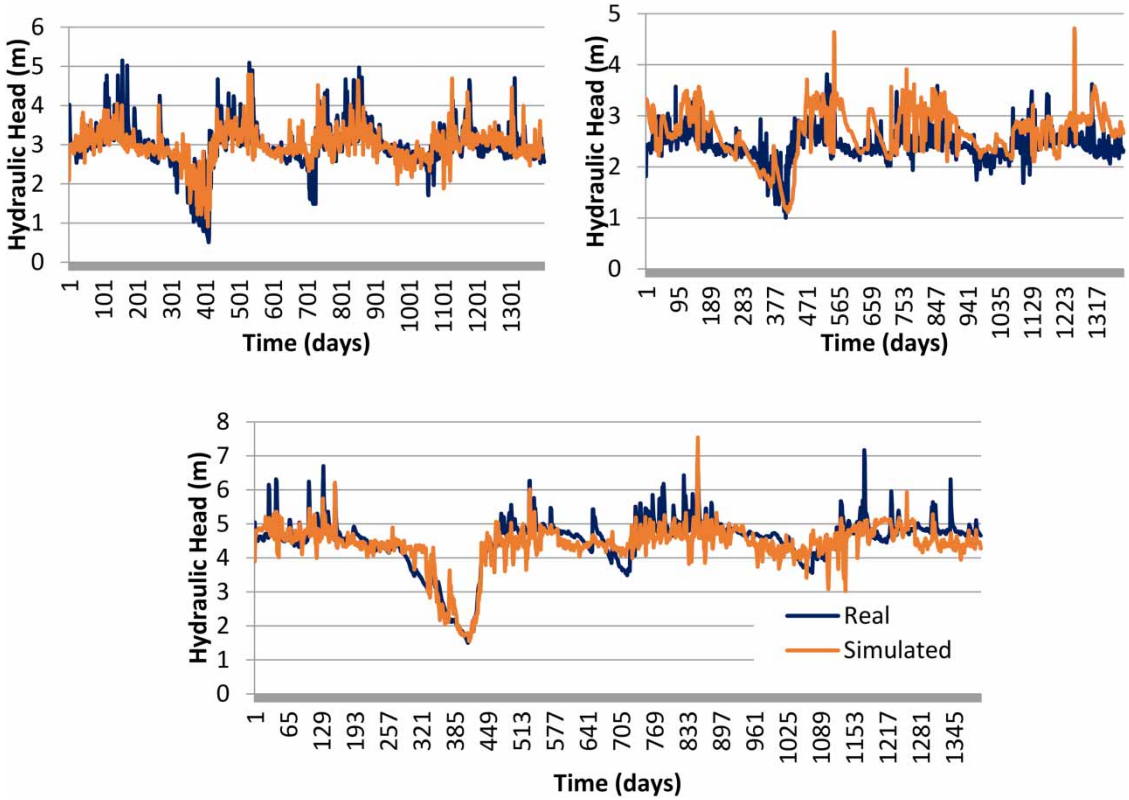

Figure 3 | Real and simulated hydraulic head for wells 11, 3 and 18.

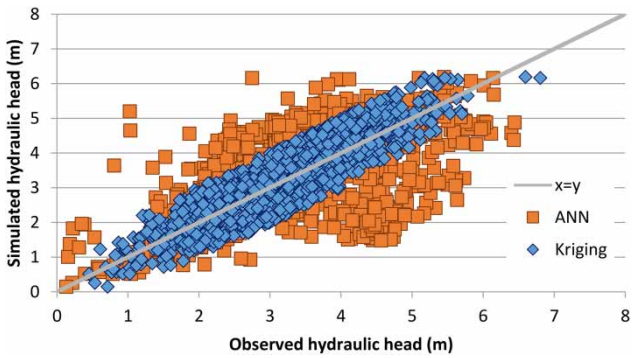

Figure 4 | Cross-validation kriging and ANN results versus observed data.

depends on the sensitivity of the output to the ANN training. In this case, well 11, albeit having an RMSE lower than well 3, ends up with slightly wider prediction intervals.

The results for the three representative wells are presented in Figure 5.

While the 90% prediction intervals were constructed using the training dataset, this does not necessarily mean that the testing data will always fall within these intervals as well. For the vast majority of available data points, especially in wells 3 and 18, the observed values of the testing dataset fall within the 90% prediction interval of the ANN, yet the total coverage is lower than the nominal value of 90%. In well 11, which has the best fit of all the other wells, only a few data points, and more specifically at some of the lowest observed values, have real values not within the prediction interval. The coverage of these intervals with respect to the observed values was, therefore, the highest for well 11 (82%) and lower for well 3 (75%) and for well 18 (79%).

ANN training model uncertainty – Monte Carlo

To evaluate the uncertainty of the ANNs training, the process described in section 'Uncertainty in ANN results' was performed 300 times, with each case having different training and testing datasets, as well as different initial random neural weights. For all wells used, the resulting box-plots of this process can be summarized in Figure 6, for the training and testing error, respectively.

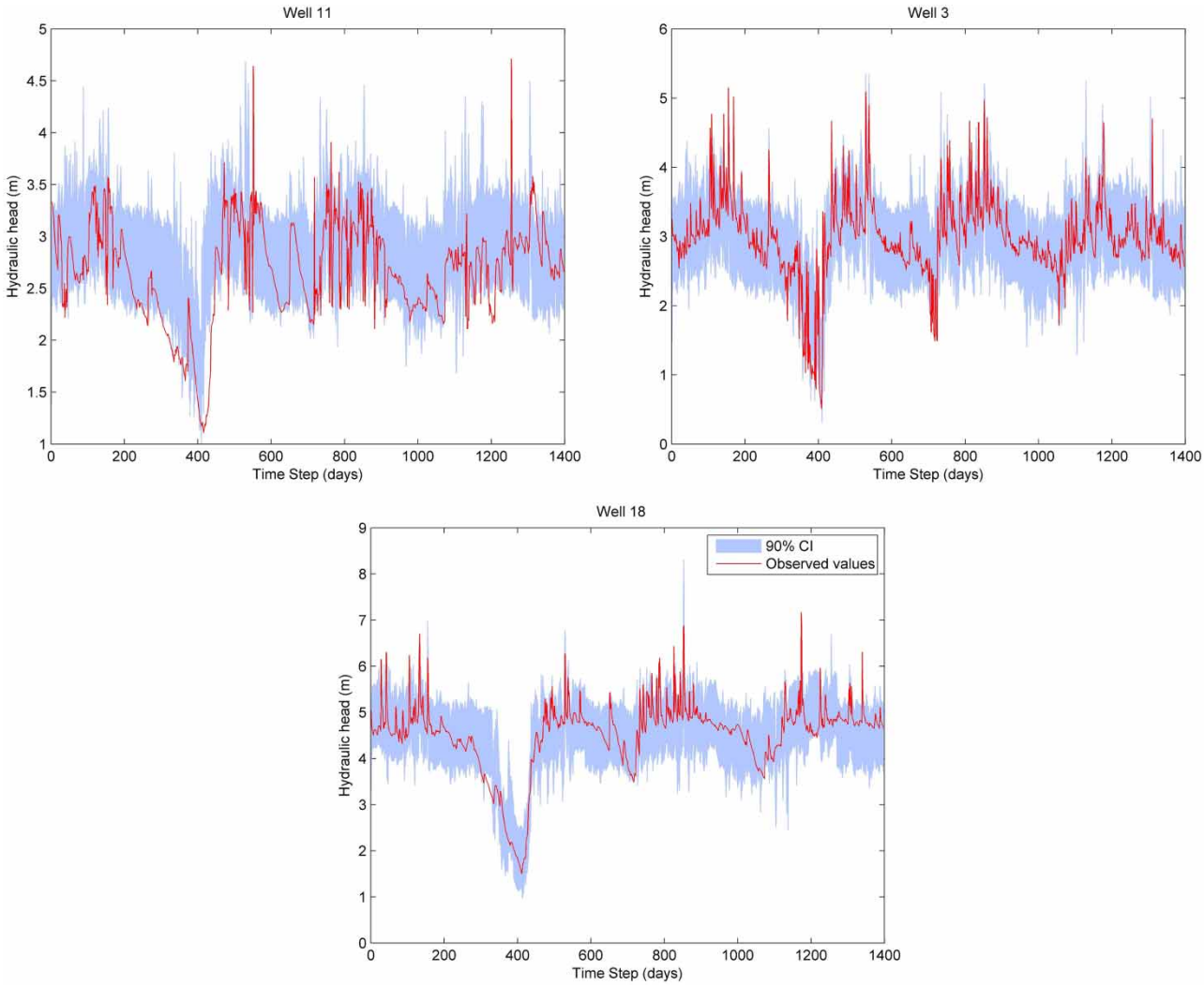

Figure 5 | 90% confidence intervals for ANN simulation.

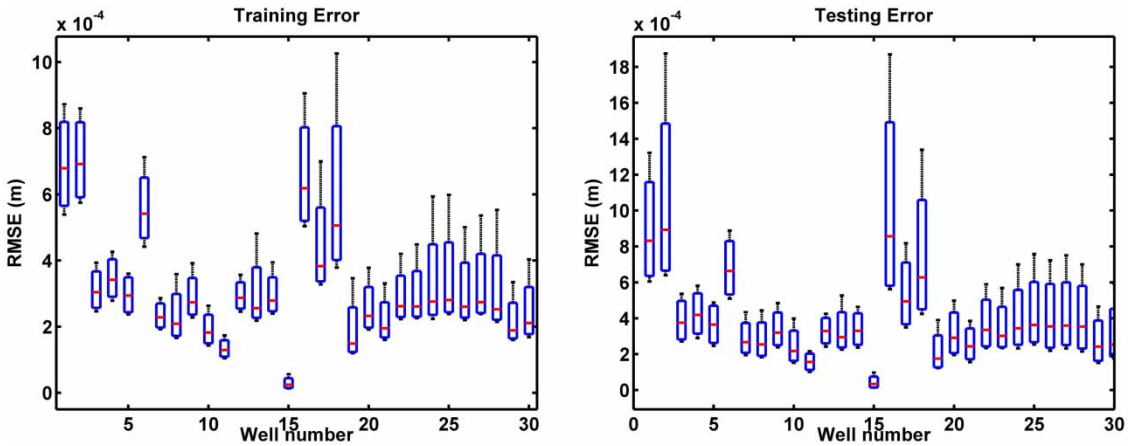

Figure 6 | Training and testing error range for training uncertainty for 30 wells.

The training error range, as expected, is smaller than the testing in most of the cases (26/30 wells). The results in terms of the representative wells are calculated and presented, together with those observed in the field values, in Figure 7.

For well 11, 94.5% of the observed values are within the 90% prediction interval. As the ANN performance deteriorates, this percentage shrinks to 90.9% for well 3 and 88.8% for the worst performance ANN category representative, well 18. The average width of the prediction intervals for each one of the representative wells are 0.153, 0.243 and 0.452 for wells 11, 3 and 18 correspondingly.

By comparing these prediction interval widths for the two different methodologies, it can be noted that the uncertainty derived for the training of the ANNs is only a small fraction of the model uncertainty calculated through the percentile methodology. This may also be attributed to the coarse nature of calculations in the percentile methodology. The second method creates narrower prediction intervals with higher coverage percentages, even for low performing ANNs, thus justifying the extra computational cost needed compared to the simple prediction confidence intervals of the previous section.

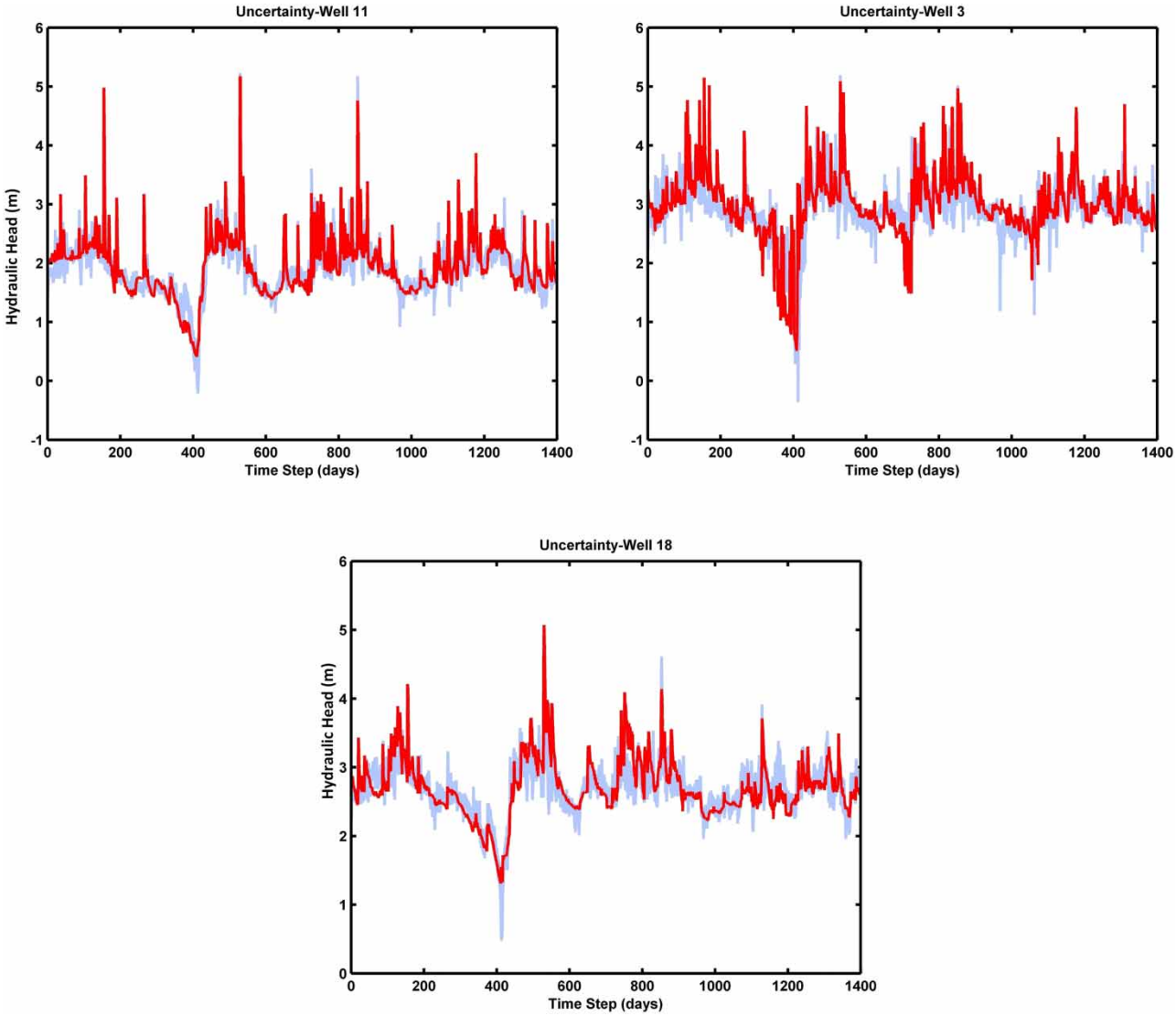

Figure 7 | Training uncertainty for three representative wells: in red, the observed in the field value and with blue, the prediction range. Please refer to the online version of this paper to see this figure in colour: http://dx.doi.org/10.2166/hydro.2019.137.

Uncertainty of the simulation algorithm due to ANN training uncertainty

For all 300 implemented trainings, the kriging part of the algorithm was also executed, resulting in a prediction interval for every studied prediction point. In Figure 8, the 90% prediction intervals for three random prediction points (namely 200, 400 and 600) are presented.

Depending on the prediction point and its proximity to observation points with good training error, the prediction intervals are either wider or narrower. More specifically, for the prediction points presented in Figure 9, the average width of the 90% prediction intervals are 0.25, 0.86 and 1.424 for prediction points 200, 400 and 600 correspondingly.

Looking at the location of these prediction points in the map, as the three prediction points fall closer, or further away from observation points with good training error, their prediction interval narrows or widens, respectively. Of the three prediction points, the one closest to observation points with good training error is the prediction point 200, while 400 lies a bit further away from the observation points. Point 600 is the one furthest away from observation data points.

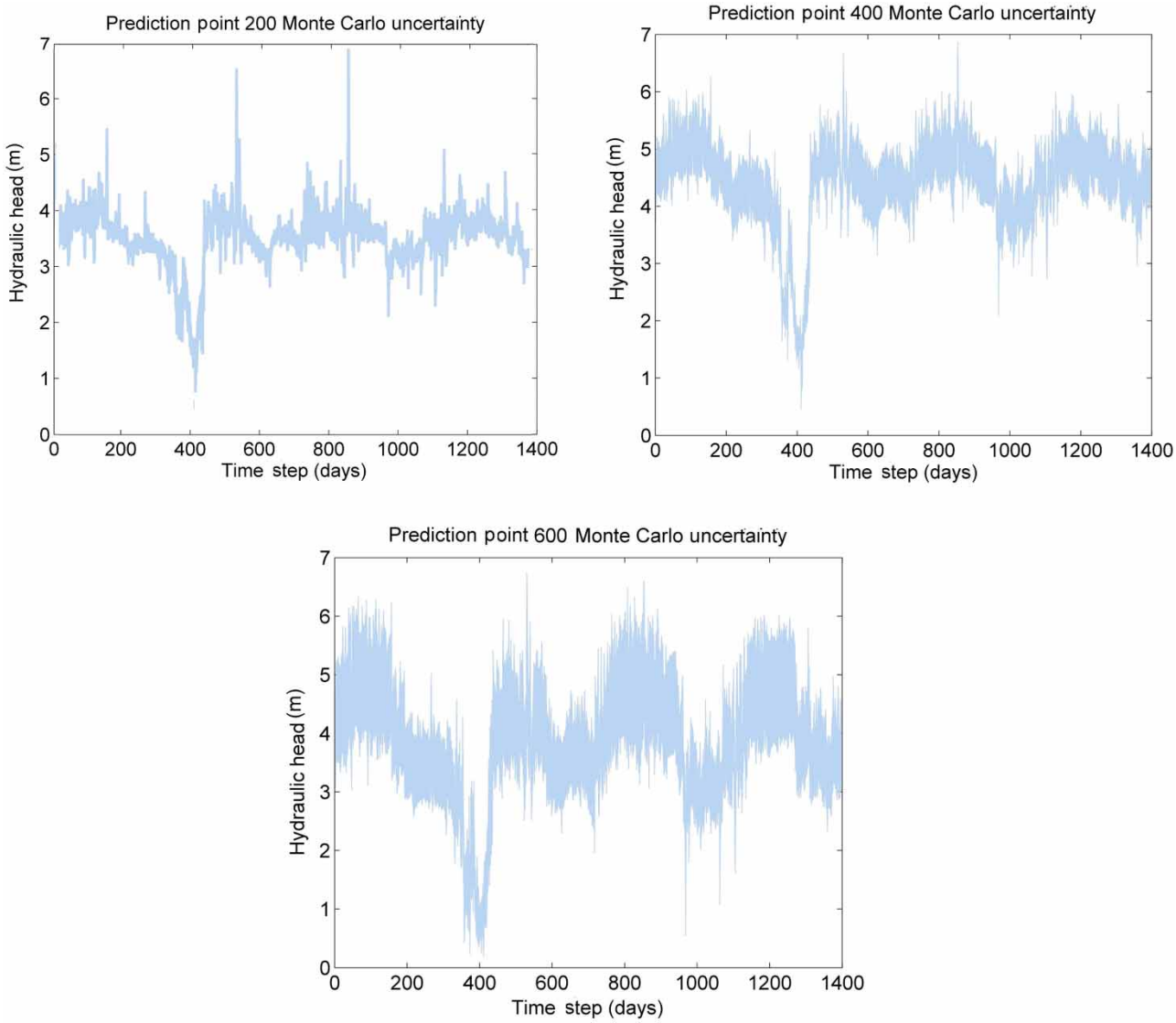

Figure 8 | Kriging 90% prediction intervals for four prediction points generated by different ANN trainings.

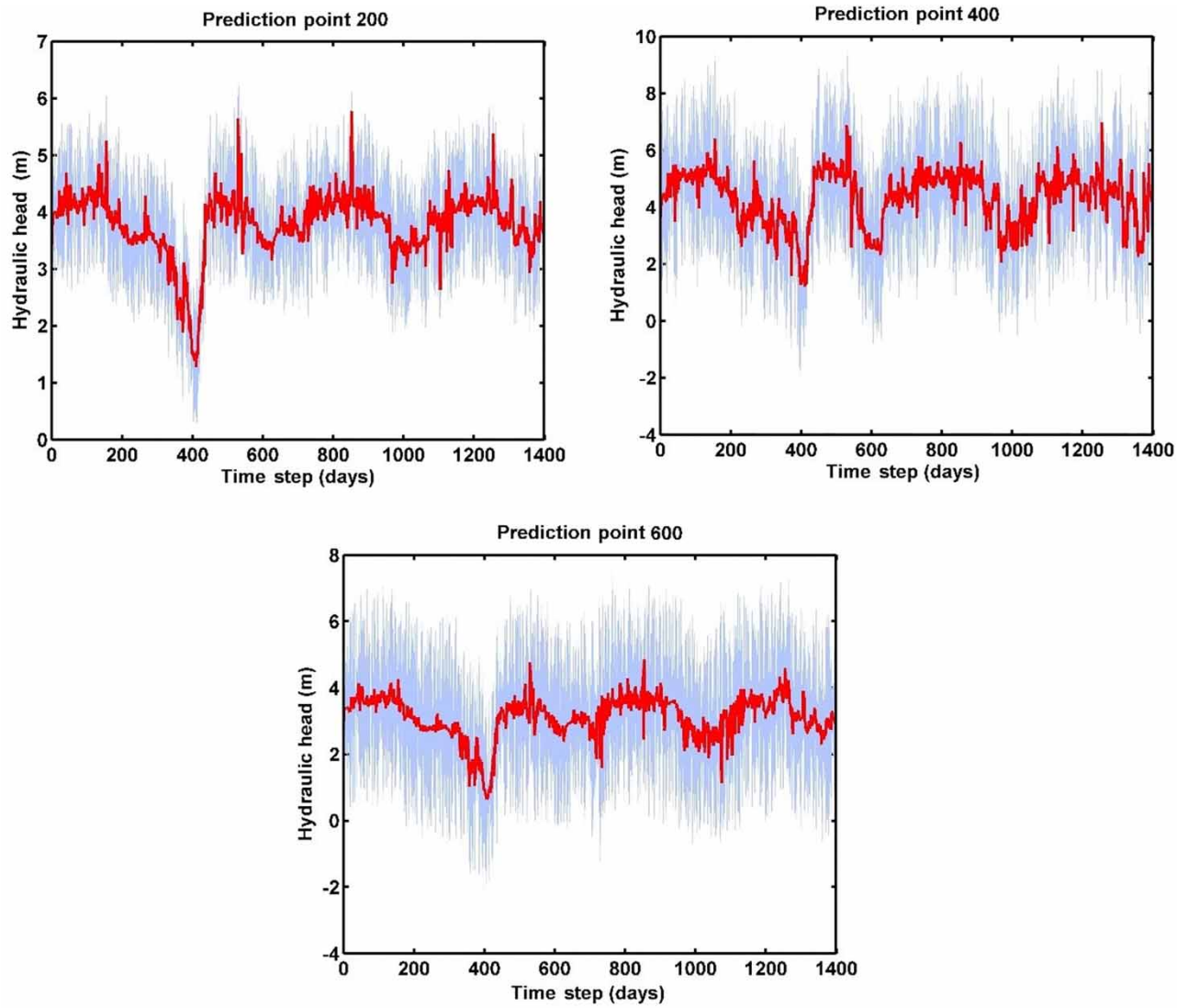

Figure 9 | Kriging parameter uncertainty for three prediction points. Please refer to the online version of this paper to see this figure in colour: http://dx.doi.org/10.2166/hydro.2019.137.

Uncertainty in hydraulic head change due to kriging parameters

Following the procedure described in the Methodologies section, the kriging parameter uncertainty is also calculated. For every prediction point and time step, 500 different data samples and hence 500 different variograms were constructed. The uncertainty was calculated for the algorithm with the use of 20 neighbors defined by the fuzzy logic system and for the exponential variogram. The 90% prediction intervals (blue range) together with the prediction derived from the initially simulated values (red lines) for the same three prediction points as before are presented in Figure 9.

In all cases, the uncertainty of the results, derived from using the described methodology, corresponding to the uncertainty due to the kriging parameter estimation is presented. The average prediction interval width for prediction points 200, 400, and 600 are 1.35 m, 2.32 m and 3.5 m correspondingly.

CONCLUSIONS

In the present study, an uncertainty analysis is performed on a combined ANN–fuzzy logic–kriging methodology for the hydraulic head simulation of an aquifer. The initial simulation

of the hydraulic head in a complex study area, in Miami, Dade County, Florida, USA had an average RMSE training error 6.63×10^{-4}m, an average RMSE testing error 7.92×10^{-4} m and cross-validation RMSE of 0.962 m. The uncertainty analysis proved that the simulation algorithm used is both consistent and accurate, especially considering the complexity of the case study and the methodology involved.

Using the percentile methodology, the ANN uncertainty can be calculated. In this case, the 90% prediction intervals are wider, compared to the ones produced by all other methodologies presented in this study. This can be attributed to the coarse nature of the calculations in this methodology.

The Monte Carlo method for assessing the uncertainty attributed to the ANN training and consecutively for sensitivity analysis of the kriging results to the ANN training is also performed. The range of training and testing error varied, without, however, having a large effect on the 90% prediction interval of the hydraulic head, in locations where data were available. The 90% prediction interval of the hydraulic head is also depicted in three prediction points, having narrow intervals in all cases.

Kriging parameter uncertainty reflects the uncertainty attributed to the observed data used. Using artificial data and the Bayesian kriging methodology, this aspect is examined. The results in terms of predicted hydraulic head intervals are close to the simulated results using the real data.

All the above-mentioned uncertainty calculations indicate that the methodology used can provide consistent and reliable results under various conditions; hence, it can be used for groundwater-level simulation, especially in complex study areas where geological information is obscure, making conventional modeling unsatisfactory.

REFERENCES

Dehghani, M., Saghafian, B., Nasiri Saleh, F., Farokhnia, A. & Noori, R. 2014 Uncertainty analysis of streamflow drought forecast using artificial neural networks and Monte-Carlo simulation. *International Journal of Climatology* **34** (4), 1169–1180.

Delbari, M., Amiri, M. & Motlagh, M. B. 2014 Assessing groundwater quality for irrigation using indicator kriging method. *Applied Water Science* **6** (4), 371–381.

Dybowski, R. & Roberts, S. 2001 Confidence intervals and prediction intervals for feed-forward neural networks. In: *Clinical Applications of Artificial Neural Networks*. Cambridge University Press, pp. 298–326.

Eslamian, S. 2014 *Handbook of Engineering Hydrology: Environmental Hydrology and Water Management*. CRC Press, Boca Raton, FL.

Fine, T. L. 1999 *Feedforward Neural Network Methodology*. Springer Science & Business Media, New York.

Fish, J. E. & Stewart, M. T. 1991 *Hydrogeology of the Surficial Aquifer System, Dade County, Florida*. US Department of the Interior, US Geological Survey, Tallahassee, FL.

Jiang, Y., Nan, Z. & Yang, S. 2013 Risk assessment of water quality using Monte Carlo simulation and artificial neural network method. *Journal of Environmental Management* **122**, 130–136.

Journel, A. G. & Huijbregts, C. 1978 *Mining Geostatistics*. Academic Press, London, pp. 40–41.

Kasiviswanathan, K. S. & Sudheer, K. P. 2013 Quantification of the predictive uncertainty of artificial neural network based river flow forecast models. *Stochastic Environmental Research Risk Assessment* **27**, 137–146.

Maiti, S. & Tiwari, R. 2014 A comparative study of artificial neural networks, Bayesian neural networks and adaptive neuro-fuzzy inference system in groundwater level prediction. *Environmental Earth Sciences* **71** (7), 3147–3160.

Pilz, J. & Spöck, G. 2008 Why do we need and how should we implement Bayesian kriging methods. *Stochastic Environmental Research Risk Assessment* **22** (5), 621–632.

Sahoo, M., Das, T., Kumari, K. & Dhar, A. 2017 Space–time forecasting of groundwater level using a hybrid soft computing model. *Hydrological Sciences Journal* **62** (4), 561–574.

Schaefli, B. & Gupta, H. V. 2007 Do Nash values have value? *Hydrological Processes* **21** (15), 2075–2080.

Shrestha, D. L. & Solomatine, D. P. 2006 Machine learning approaches for estimation of prediction interval for the model output. *Neural Networks* **19** (2), 225–235.

Tapoglou, E., Karatzas, G. P., Trichakis, I. C. & Varouchakis, E. A. 2014 A spatio-temporal hybrid neural network-Kriging model for groundwater level simulation. *Journal of Hydrology* **519** (Part D), 3193–3203.

Tapoglou, E., Varouchakis, E. A. & Karatzas, G. P. 2018 Uncertainty estimations in different components of a hybrid ANN–Fuzzy–Kriging model for water table level simulation. In: *HIC 2018, 13th International Conference on Hydroinformatics*, Palermo. Vol. 3, pp. 2042–2050.

Trichakis, I., Nikolos, I. & Karatzas, G. P. 2011 Comparison of bootstrap confidence intervals for an ANN model of a karstic aquifer response. *Hydrological Processes* **25** (18), 2827–2836.

USGS 2014 *Geological Units in Miami-Dade county, Florida*. mrdata.usgs.gov/geology/state/fips-unit.php?code=f12086 (accessed 7 March 2016).

First received 11 December 2018; accepted in revised form 9 April 2019. Available online 29 April 2019