Semi-seasonal groundwater forecast using multiple data-driven models in an irrigated cropland

Alessandro Amaranto, Francisco Munoz-Arriola, Gerald Corzo, Dimitri P. Solomatine and George Meyer

ABSTRACT

In agricultural areas where groundwater is the main water supply for irrigation, forecasts of the water table are key to optimal water management. However, water management can be constrained by semiseasonal to seasonal forecast. The objective is to create an ensemble of water table one- to five-month lead-time forecasts based on multiple data-driven models (DDMs). We hypothesize that data-driven modeling capabilities (e.g., random forests, support vector machines, artificial neural networks (ANNs), extreme learning machines, and genetic programming) will improve naïve and autoregressive simulations of groundwater tables. An input variable selection method was used to carry the maximum information in the input (precipitation, crop water demand, changes in groundwater table, snowmelt, and evapotranspiration) and output relationship for the DDMs. Five DDMs were implemented to forecast groundwater tables in an unconfined aquifer in the Northern High Plains (Nebraska, USA). Root mean squared error (RMSE) and Nash–Sutcliffe efficiency index were used to evaluate the skill of the model and three hydrologic regimes were determined statistically as high, mid, and low groundwater table levels. Additionally, varying storage regimes were used to construct rising and falling limbs in the tested well. Results show that all models outperform the baseline for all the lead times, ANNs being the best of all.

Key words | data-driven models, ensemble, groundwater, semi-seasonal forecast

Alessandro Amaranto
Francisco Munoz-Arriola (corresponding author)
George Meyer
Biological Systems Engineering Department,
University of Nebraska-Lincoln,
Lincoln, NE 68588,
USA
E-mail: *fmunoz@unl.edu*

Alessandro Amaranto
Gerald Corzo
Dimitri P. Solomatine
IHE Delft Institute for Water Education,
Delft,
The Netherlands

Dimitri P. Solomatine
Water Resources Section,
Delft University of Technology,
Delft,
The Netherlands
and
Water Problems Institute of RAS,
Moscow,
Russia

ABBREVIATIONS

DDM	Data-driven model
GW	Groundwater
ANN	Artificial neural network
P	Precipitation
ET	Evapotranspiration
SNM	Snowmelt
AD	Average water demand
IVS	Input variable selection
CIVS	Constrained input variable selection
EnsAnn	Ensemble artificial neural network
EnsSvm	Ensemble support vector machines
EnsRf	Ensemble random forests
EnsElm	Ensemble extreme learning machines
EnsGP	Ensemble genetic programming
RMSE	Root mean squared error
NSE	Nash–Sutcliffe efficiency index
ROC	Receiver operating characteristic curve
MLP	Multilayer perceptron
SVM	Support vector machines
RF	Random forest
GP	Genetic programming
ELM	Extreme learning machines
AR	Autoregressive model
LWR	Low water range

doi: 10.2166/hydro.2018.002

MWR Middle water range
HWR High water range
AUC Area under the curve
RL Rising limb
FL Falling limb

INTRODUCTION

Groundwater is a key resource to sustaining hydrological conditions of a watershed as well as agricultural activities. The availability of groundwater (GW) during droughts helps when it is doubtful there will be enough precipitation to sustain crops and their yields. When withdrawals exceed the recharge rate of an aquifer for a long period, GW depletion occurs. Common consequences of aquifer overexploitation are water rationing, drying up of wells, less water in streams and lakes, water-quality degradation, increased pumping costs, land subsidence, decreased well yields (Bartolino & Cunningham 2003; Nayak *et al.* 2006), and unsustainable agriculture. Therefore, implementing effective (and, preferably, efficient) water management strategies is crucial for conserving hydrological conditions and sustaining water and agricultural resources and ecosystem services.

Reliable water supply policies require accurate estimations of current and future water table depths and their fluctuations (see, for example, Coulibaly *et al.* 2001). For this purpose, physically based, statistical, and data-driven modeling techniques are widely used. For example, Hanson *et al.* (2010) implemented the physically based model MODFLOW and the farm process package (MF-FMP), parameterizing the micro- and macro-scale crop irrigation requirements or evapotranspirative needs to simulate the conjunctive use of surface water and groundwater. Also, models like MF-FMP are promising tools to reproduce supply-constrained and demand-driven hydrologic budgets; however, their implementation requires a full hydrogeological characterization of the aquifer including anisotropic and spatially distributed properties (Coppola *et al.* 2003a; Mohanty *et al.* 2010). Places where networks of groundwater well measurements enable local to regional water management, can be used to track spatiotemporal variability of groundwater in response to urban as well as irrigation water supplies and demands (Scanlon *et al.* 2012). In agricultural areas where GW-based irrigation is used to satisfy evapotranspirative needs, variables such as evapotranspiration, crop water demand, precipitation and groundwater well levels could integrate the imbedded complexity of aquifer recharge with respect to streamflow, precipitation fluctuations, and well management.

Developments in the area of machine learning have greatly expanded the capabilities of data-driven models (DDMs) to diagnose and forecast hydrological states (Maier & Dandy 2000; Solomatine *et al.* 2009). DDMs are recognized by their ability to reconstruct the relationships among inputs, states and outputs of a system, without an explicit knowledge of the system's physical behavior. For this reason, DDMs can play a complementary role to physically based models and help overcome some of the limitations associated with multiple and complex variables (Coppola *et al.* 2003b).

DDMs are widely used in hydrology (Abrahart *et al.* 2012) for applications ranging from rainfall–runoff modeling (Solomatine & Dulal 2003), river flow forecasting (Dibike & Solomatine 2000; Akhtar *et al.* 2009; Taormina & Chau 2015a), flood forecasting (Campolo *et al.* 1999; Solomatine & Xue 2004) to drought forecasting (Kim & Valdés 2003; Le *et al.* 2016). The most widely used technique is artificial neural networks (ANNs). However, the number of studies using DDMs to forecast water table levels is limited. For example, Daliakopoulos *et al.* (2005) tested multiple ANN architectures to identify the most suitable to predict water table fluctuations over an 18-month lead time. Coppola *et al.* (2005) studied the ability of ANNs to predict water table levels with lead time of 30 days near a public supply wellfield. Sun *et al.* (2016) analyzed the ability of ANNs to predict water table depth in a swamp forest in Singapore. They concluded that accurate estimates could be obtained with a daily lead time, whereas the performance decreased when the lead time was increased to a week. Tsanis *et al.* (2008), Mohanty *et al.* (2015), and Djurovic *et al.* (2015) also used ANNs to forecast the GW levels at weekly and monthly lead times, discovering that additional input variables, such as precipitation, evaporation, and river stage, increased the model performance in single or multiple wells. Regarding semi-seasonal to seasonal forecasts at six months' lead time, Varouchakis (2016) and Lohani & Krishan (2015)

used a Kalman filter adaptation algorithm with exogenous inputs and ANNs in the Mires basin in Crete, Greece, and in Punjab, India, obtaining accurate estimations of GW levels.

The need for better planning and management of water resources in the world has driven the focus on increasing lead times and on determining how the uncertainty associated with increased lead times could influence results. Also, no clear procedure exists for selecting appropriate DDM to represent GW dynamics. Optimal model tuning has evolved, and one of the main recommendations by Galelli *et al.* (2014), for example, is to improve input variable selection (IVS) for a better performance.

The human impact on groundwater is quite high in certain regions and, therefore, is an important variable to include, even though it is not easily measured. For a proper assessment of forecast capabilities, DDMs have been studied per regimes in surface water systems (Corzo & Solomatine 2007). These regime analyses in surface hydrology aim to evaluate models' performance in the rising and falling limbs (FLs), as well as in high- and low-flow conditions. In GW systems, it could be used to quantify and improve GW predictability groundwater withdrawals and recharge (falling and rising limb, respectively), as well as in water shortages and surpluses (low and high water table level, respectively).

Therefore, the objective of this paper is to assess and compare the capability of five DDMs to forecast water table levels one to five months ahead in response to integrated hydro-climatological forcings and water management. This will be achieved by developing a fully fledged framework employing IVS (including lags) and assessing the DDMs' performance for different regimes of GW levels.

The hypotheses of the present work are: (1) the models' performance will improve when crop water demand is included in the input set. This hypothesis is based on the assumption that crop water demand is a valid proxy to represent the influence of human intervention (when pumping data are unavailable); (2) variability of the performance of DDMs can be significantly different for different groundwater levels' regimes; and (3) models' performance vary with respect to the increase in lead time. Thus, the scientific question for this study is to find out the extent of such variation.

The proposed approach complements previous studies by performing a comparative analysis of DDMs for forecasting GW levels at five different lead times. For this experimental analysis, we tested an improved version of the exhaustive search IVS technique, which allowed us to have a better variable input set. In addition, the human impact on the GW system was addressed (in the absence of pumping data) by introducing the crop water demand as a proxy. This choice relies on the assumption that the amount of water pumped and the available surface water are proportional to the evapotranspirative needs of crops (also known as crop irrigation requirements). Moreover, this paper assesses the performance of different models in different hydrological regimes by quantifying their predicting capabilities in the falling (withdrawal) and rising (recharge) limbs of the water level, as well as in water shortage and high water availability conditions.

The paper is structured as follows: the study area and the data available are described and a basic characterization of the groundwater system is provided. Then, methodology is presented, including the IVS methods, the modeling techniques, and the metric of performance. Results and discussions follow. The last section of the paper summarizes the main findings and conclusions of the study, as well as its limitations and future directions.

STUDY AREA AND AVAILABLE DATA

The study area is located in the central-east part of the state of Nebraska, USA, in the Middle Platte-Prairie hydrogeological unit (Figure 1). This area is crossed by the Platte River, which brings water from the Rocky Mountains to the Missouri River, draining northeast Colorado, southeast Wyoming, and central Nebraska (Eschner *et al.* 1983). The Middle Platte-Prairie is part of the High Plains aquifer (Figure 1(a)), which is the largest (450,000 km^2) in the United States and has been intensively developed for irrigation purposes. The High Plains aquifer water levels have shown declines of more than 30 m in the past 30 years, leading to a reduction in saturated aquifer thickness in some areas of more than 50% (Scanlon *et al.* 2012). In the Nebraska portion of the High Plains aquifer (Figure 1(b)), the number of registered wells has grown from 1,200 in

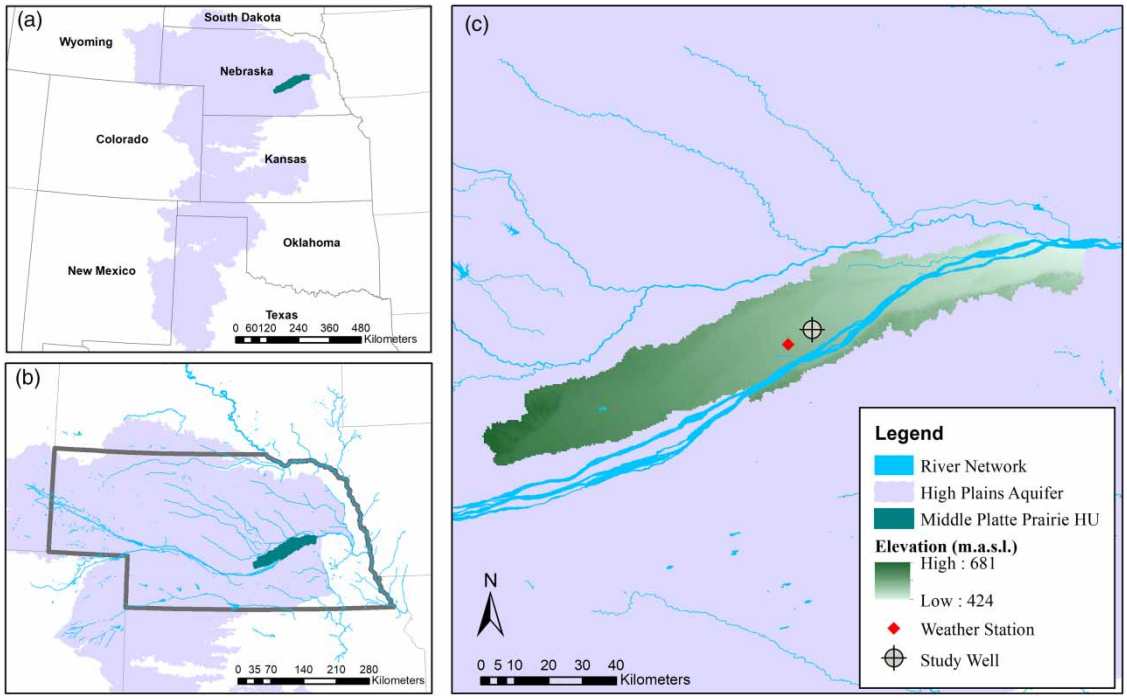

Figure 1 | Location of the High Plains aquifer (a), of the State of Nebraska (b), and of Middle Platte-Prairie hydrogeological unit (c).

1936 to about 100,000, serving about 85% of the state's irrigation land (Flowerday *et al.* 1998; Wen & Chen 2006). According to Scanlon *et al.* (2012), groundwater storage decreased about 330 km^3 between the 1950s and 2007. This amount represents about 8% of the available GW storage. Depletion is particularly severe in the central and southern part of the aquifer (Kansas and Texas). Those areas in particular can be considered a hot spot for aquifer depletion.

The High Plains aquifer consists of hydraulically connected deposits of late Tertiary and Quaternary age. Among those, the Late Tertiary age Ogallala formation (a heterogeneous deposit of interlayered stream sediments, lakebeds, and eolian sand, silt, and clay) covers about 342,000 km^2 of the aquifer (about 75% of the total area).

The Middle Platte-Prairie hydrogeological unit (Figure 1(c)) has an area of approximately 2,800 km^2. It is constituted by unconsolidated Quaternary alluvial deposits (generally more permeable than those of the Ogallala formation; Gutentag *et al.* 1984). The average saturated thickness of the area is about 100 m. The main use of the land is corn agriculture. NAS-USDA (2011) estimated that the irrigated corn yield was between 2 and 2.5 tons/ha in 2011 in this particular area (state average = 2.2 tons/ha). The irrigation system is usually a center pivot sprinkler. This type of system has several pipes joined together to form a robotic arm mounted on a tower. The robotic arms move in a circular pattern and release water through sprinklers. Water is usually pumped from the aquifer and fed to the sprinklers from the pivot point located at the center of the circle.

Precipitation (P, mm) data in the Middle Platte Prairie were monitored by a weather station from the High Plains Regional Climate Center located in Central City. Hydroclimatic variables observed by the GLDAS (Global Land Data Assimilation System; Rodell *et al.* 2004) provided monthly estimations of evapotranspiration (ET, mm) and snowmelt (SNM, mm) at a spatial resolution of (1/8)° latitude × longitude (≈16 km). Pumping data were unavailable for this particular region. As a consequence, pumping is partially represented by the use of the corn water demand as an input variable. This choice relies on the assumption that the amount of water pumped by farmers is proportional to the amount of water required by the crop. Average water demand data (AD, mm) were obtained by Kranz *et al.* (2008), who computed the long-term monthly average of

water amounts required by corn. It is, therefore, a time series that repeats identically every 12 months and does not take into account inter-annual demand variation.

The USGS (2015) has monitored water table data for the Middle Platte-Prairie hydrogeological unit. Data are available daily and monthly. Figure 2(a) shows the GW time series for the monitoring well in the hydrogeological unit, while Figure 2(c) zooms into the three years from January 2012 to December 2014. Figure 2(c) shows major water level declines during the crop growing season, followed by partial recharge during the off-season. Water level declines are due to withdrawal (pumping) from farmers and to the high evapotranspiration (Figure 2(b)) rate characterizing the crop growing season. The highest water demand for corn is in July, when the crop grows tassels. Recharge to the aquifer occurs when withdrawal stops (September–October) and when there are precipitation and snowmelt (Figure 2(b)). Snowmelt usually occurs at the end of winter and the beginning of spring (February–April). Precipitation is generally high during autumn and spring, with sporadic events during winter and summer. Soil freezing during winter prevents rainfall from recharging the aquifer.

The water year is characterized by strong water level depletion from April to September–October, followed by a fast natural recovery from October to November. The recovery usually stops during winter and starts again

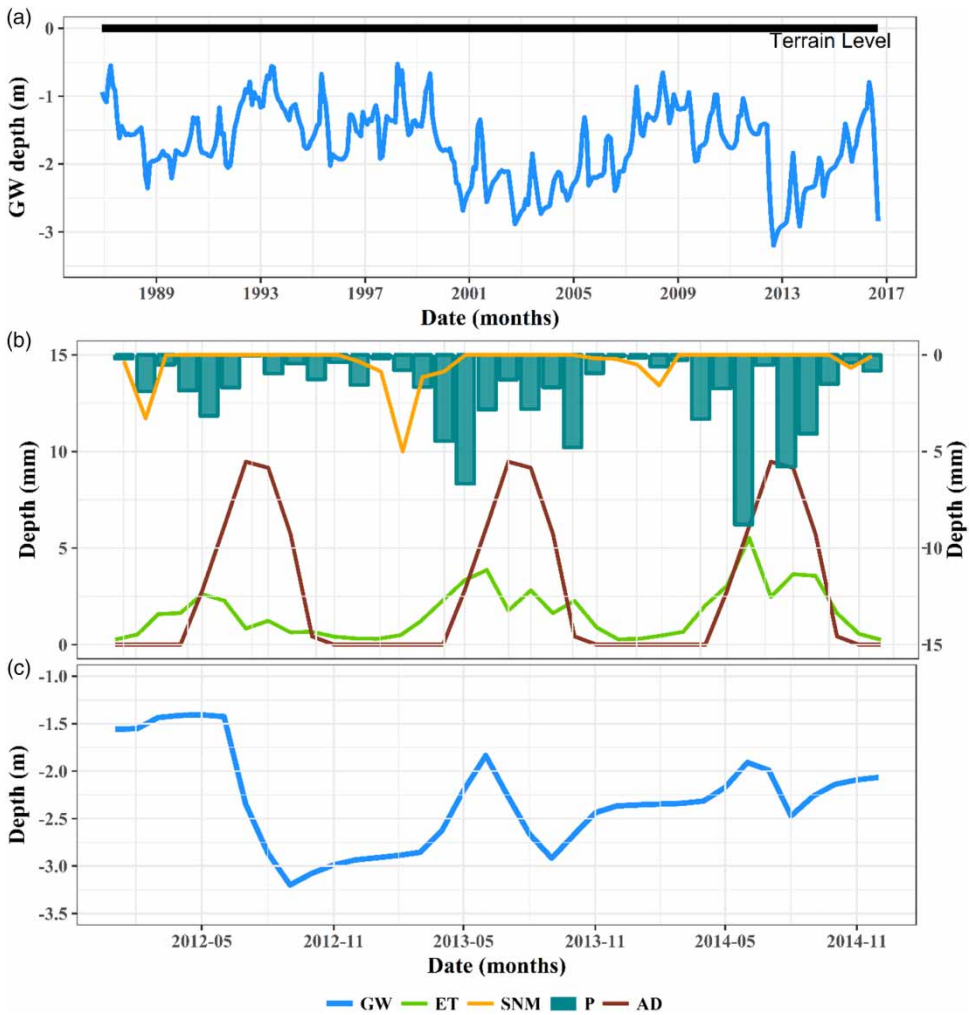

Figure 2 | (a) Time series of groundwater (GW) levels in the monitoring well (USGS 2015) for 1987 to 2016; (b) precipitation (P), evapotranspiration (ET), snowmelt (SNM), and corn water demand (AD) from 2012 to 2014; (c) GW level from 2012 to 2014.

with snowmelt and precipitation at the beginning of spring.

As can be seen from Figure 2(a), water level declines were particularly severe during the July 2012–February 2013 period, when the lowest amount of precipitation ever recorded in Nebraska led to an increase in withdrawals from the aquifer.

Table 1 provides a summary of the descriptive statistics of the data, compared with the average value in the Northern High Plains (where available).

METHODOLOGY

Methodological framework

To test the hypotheses above, a suite of data-driven modeling approaches were used to forecast GW levels in response to environmental forcings and anthropogenic regulations (Figure 3). The implementation of this multi data-driven model testing starts by dividing the data into training and testing sets. Data were also normalized (block data division and transformation). To select the most relevant input variables and lags, the so-called constrained input variable selection (CIVS) procedure is developed and implemented. CIVS is applied to precipitation, snowmelt, evapotranspiration, GW level, and crop water demand in the training set. The resulting variables from CIVS are then used to force the models.

The training set is split into training and cross-validation set. The training set is used to optimize the value of the model's parameters (e.g., the weights in the ANN). The cross-validation set is used to optimize the architectures of the DDMs (e.g., hidden nodes number in ANN), aiming at minimizing the error on this set. The procedure is repeated ten times (ten-fold cross validation) for each of the five models, generating an ensemble of ten neural networks (EnsAnn), ten support vector machines (EnsSvm), ten random forests (EnsRf), ten genetic programming models (EnsGp), and ten extreme learning machines (EnsElm). Simple ensemble averaging of each model type outputs is used. The testing set is used to evaluate and compare the performances (root mean squared error (RMSE) and Nash–Sutcliffe efficiency index (NSE)) of the aforementioned ensemble models. Performances of models are also evaluated at different water level regimes. Discriminating ability of the models is tested per each water regime by means of the receiver operating characteristic (ROC) curve.

Data division and transformation

Theory of DDMs states that the training and test sets come from approximately the same statistical distribution. To achieve this, several splits of data can be made and then the training and testing sets in each split compared. Consequently, the split providing the closest statistical properties of the training and testing sets can be chosen. However, particular implementations of DDMs are often constrained by requirements relevant to a particular application area, and in the hydrological field, it is often required to test the predictor on the most recent data. As a consequence, the authors decided to select the initial 70% of the data (February 1987–October 2007) as the training set, and the remaining 30% (November 2007–September 2016) as the testing set. The training set was then normalized, and the normalization parameters (minimum and maximum) were used to normalize the testing set.

Table 1 | Descriptive statistics of the dataset for the period February 1987–September 2016 (356 instances) in the study area (SA), compared with the average value in the Northern High Plains (NHP)

	GW depth (m) SA	P (mm/mo) SA; NHP	SNM (mm/mo) SA; NHP	ET (mm/mo) SA; NHP	AD (mm/mo) SA
Minimum	−3.2	0; 1.18	0; 0	0.04; 1.38	0
Maximum	−0.52	263.3; 195.2	70.4; 43.06	175.1; 134.43	248.2
Average	−1.71	61.8; 51.0	3.23; 2.72	57.9; 48.42	59.6
St. deviation	0.54	52.9; 39.0	8.44; 5.19	45.7; 36.17	87.4
Skewness	−0.23	1.09; 0.93	4.71; 3.47	0.60; 0.38	0.98

GW, groundwater; P, precipitation; SNM, snowmelt; ET, evapotranspiration; AD, corn water demand.

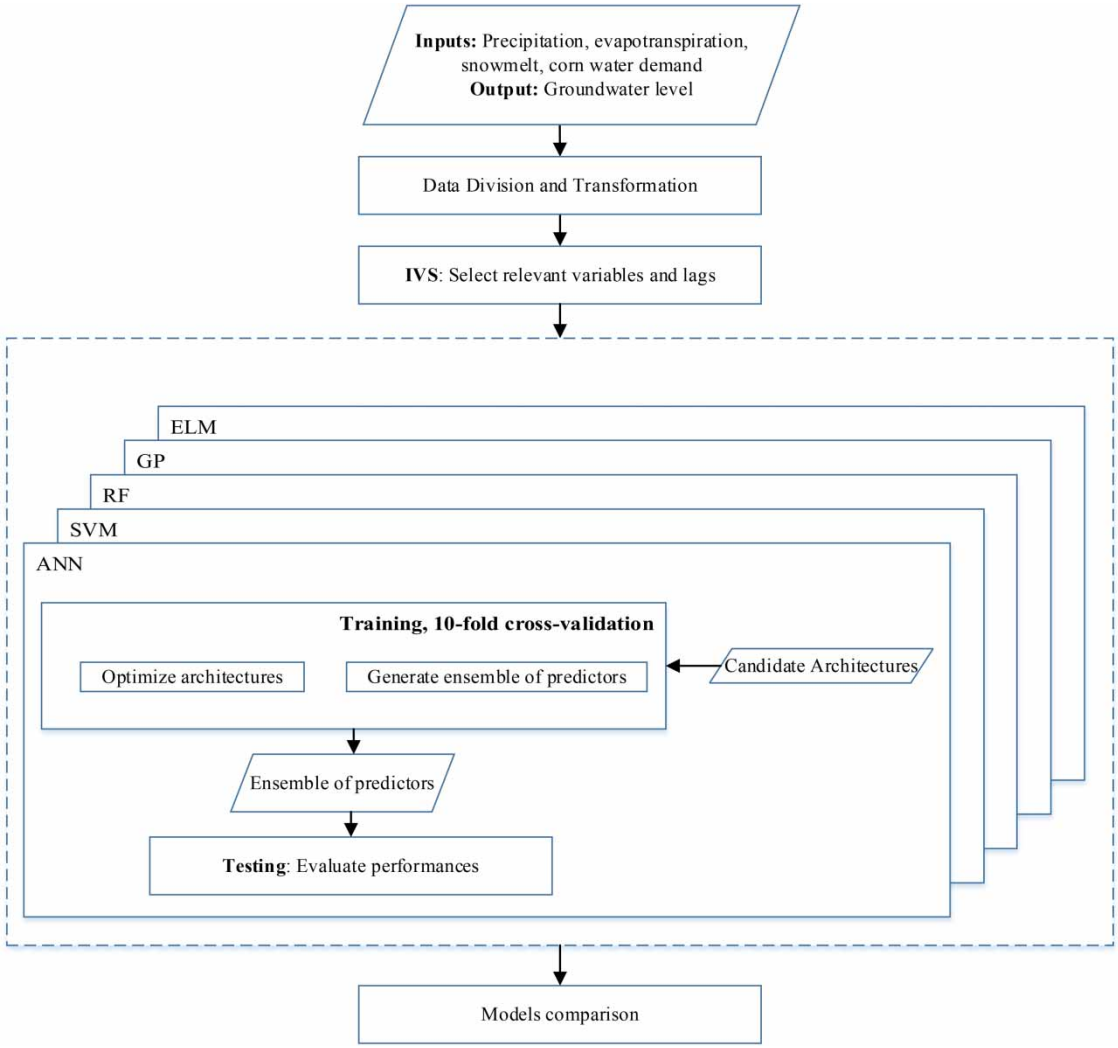

Figure 3 | Methodological framework.

Input variables selection

One of the most critical aspects in the implementation of DDMs is deciding about the list of the input variables, a procedure often referred to as 'input variables selection' (IVS). Ideally, if the dynamics of the system are clearly understood, the input variables (input vector space) could be chosen by a domain expert. When the knowledge of the system is insufficient, one option would be to perform an exhaustive search in the input space to select the best input variables subset. The input selection process can be computationally expensive, for example, to select up to n input variables out of n, there are 2^n-1 possible combinations. The complexity of the problem grows further when considering also the appropriate lags to be chosen. Therefore, finding the best input subset for a DDM requires an algorithm which is also computationally efficient. This can be posed as an optimization problem, e.g., Bowden *et al.* (2005) proposed an IVS method based on a combination of a genetic algorithm and general regression neural network (GAGRNN), May *et al.* (2008) and Elshorbagy *et al.* (2010a) suggested using a partial mutual information based selection algorithm, and Galelli & Castelletti (2013) reported a tree-based iterative search method. Interested readers can find an evaluation framework of IVS algorithms in Galelli *et al.* (2014). For this study, we use exhaustive search which is however

constrained by rules based on the knowledge of the GW physics. This method is referred to as the 'constrained input variable selection' (CIVS) method, and it is implemented as follows.

The basic environmental variables included in the development of the models are P, ET, SNM, AD, and GW level. Performing an exhaustive search on those non-lagged variables would suggest considering $d = 2^5 = 32$ candidate input sets. When we allow for k possible lags for each of the variables, the number of candidate sets increases to d^k. To limit the complexity, we have formulated a number of constraining rules aimed at reducing the number of candidate input sets and which were made part of the CIVS algorithm. The following five rules were chosen for implementation:

1. The number of lags for the autoregressive GW term is fixed to two, and the lags corresponds to the two most recent observations.

This choice relies on the fact that the autocorrelation of GW level decreases as the lag increases. Therefore, the most recent observations can intuitively be considered the best predictors. Considering that autoregressive models' performances did not improve when the order of the model was increased beyond two, the maximum number of lags was fixed accordingly (even though there is no unique way to determine the number of lags to be used as predictors).

2. The maximum number of lags for P, ET, and SNM is equal to 3. Being x any of the three aforementioned variables, the only lagged variables included are x_{t-1}, x_t, and x_{t+1} (under the assumption of perfect monthly forecast).

The case study under investigation is a shallow unconfined aquifer. The water table depth varies between 0.5 and 3.5 meters below the soil surface. Consequently, the effect of the meteorological variables such as P, ET, and SNM on water table changes occurs in a relatively short period of time. Considering this, and after time series examinations on the inputs and output variables, the search in the input space was limited to the lags corresponding to $t-1$, t, and $t+1$.

3. Among the three aforementioned variables, at least two must be considered in the input set at the same time.

The study area is in an intensely cultivated region (high ET influence) in Nebraska. It is also located in the area of the aquifer where the precipitation and snowfall are higher. It is, therefore, likely that at least two of the three variables play a role in the dynamics of the GW level changes.

4. Lag 'jumps' are not allowed. This means that if x_{t+1} is considered as an input, then x_{t-1} cannot be an input candidate in the considered subset to predict the output y.

This choice is based on the reasonable assumption that if x_{t+1} is considered a driver for changes in y, then the only other reasonable driver would be x_t, rather than x_{t+1}.

5. The choice of including AD or not is a binary variable: if m is the number of candidate subsets resulting from implementing rules 1 to 4, the overall number of candidates will be $2\,m$. Half of them will be the original candidates, the other half will be the same candidates with the addition of the crop water demand in the time interval $(t,\ t+\Delta t)$, where Δt is the forecasting lead time (one to five months).

Since AD does not change from year to year, future estimations of the variable are available. This rule represents the choice of including or not AD in the input set.

After implementation of the rules, the total number of input subsets was reduced from 32,768 (if we consider five variables and three lags) to 346. Then, the CIVS algorithm was run through each of the 346 possible combinations of the (lagged) variables. For each combination, the data were divided into a training and a testing set; a ten-fold cross-validation was performed on the training set to train an ensemble of ten multilayer perceptrons; the average RMSE on the cross-validation set was computed; and the result was stored. The best input subset was the one that minimized the average RMSE on the cross-validation set. The effectiveness of the CIVS method was assessed by comparing the performance of the models with those obtained by implementing the GAGRNN algorithm developed by Bowden *et al.* (2005). The effect of the perfect forecast assumption is instead assessed by comparing the CIVS results with those obtained with no input at $t+1$ (no perfect forecast).

Modeling techniques

The selected variables from the CIVS method were the input for the five DDMs implemented in this study. The five models had a standardized and consistent training set used to produce GW level forecasts from one- to five-month lead times. The techniques used are nonlinear statistical (learning) models whose characterization is provided below and in the Abbreviations.

Artificial neural networks

Multilayer perceptron (MLP) neural networks are a widely used and very well-developed technique (Haykin 1999), and they are indeed also widely used in water-related studies (see, for example, Elshorbagy *et al.* 2010a; Abrahart *et al.* 2012). A MLP is constituted by an input, a hidden, and an output layer. The input layer has as many nodes as the number of inputs. The number of nodes in the hidden layer is usually proportional to the complexity of the problem analyzed. The output layer usually has a single node. The connection between layers is ensured by a set of weights, which express the strength of the connection. Nonlinearity is provided by a sigmoidal transfer function in the hidden layer.

The structure of an ANN may be subject to optimization, and it is typically optimized in establishing the number of nodes in the hidden layer. The number of neurons in the hidden layer in each member of the ensemble was optimized from a set of values ranging from 3 to 15. The resilient backpropagation algorithm was used to train all neural networks using the R package *RSNNS* (Bergmeir & Benítez 2012). After iterative experimentation with the learning function parameters values, it was found that their choice was not affecting the speed of the convergence.

Support vector machines

Support vector machines have their foundation in the pioneering works of Vapnik (e.g., Vapnik 1998, 2013). They were originally developed for classification problems and attracted a great deal of attention because of their peculiarity of being a linear machine with the capability of implementing nonlinear class boundary. For that to be possible, they map the input space into a higher dimensional space, where it is possible to find a set of linear models (hyperplane) that maximizes the classification accuracy. Among those, the best one will be the one that maximizes the separation between classes. The instances located closer to the hyperplane margin are called 'support vectors'. Regression support vector machines (SVM) have been developed from the idea of producing a model that can be applied to nonlinear problems using few support vectors (Witten & Frank 2005).

There are several parameters in SVM to be identified by optimization. They are typically named C, ε, and γ. C determines the tradeoff between the flatness of the regression and the magnitude of the error, ε determines the maximum allowable error, and γ is the kernel parameter. The values of C, γ, and ε for each member of the ensemble were optimized from the set of values summarized in Table 2. The R package *e1071* (Dimitriadou *et al.* 2009) was used to train the SVM models with radial basis function kernel.

Random forests

Random forests (RF) is a relatively new machine learning technique that belongs to the area of 'ensemble learning' (EL). Like all other EL methods, RF generates a series of linear predictors and then aggregates the results of each predictor. Each predictor in the ensemble is created using a random selection in the input space at each node of the tree. This randomness has been shown to perform well in both classification and regression problems, and it ensures robustness against overfitting and outliers (Breiman 2001).

The number of trees (*NT*) and the number of variables (*NV*) to be used at each split are the parameters to be found during training. They were optimized from a set of NT values ranging from 1 to 5 and from NV ranging from

Table 2 | Minimum, maximum, and sequence type for the candidates' tuning parameters used to optimize the SVM ensemble architecture

	C	γ	ε
Minimum	0.0313	0.01565	0.0005
Maximum	16	2	1
Sequence type	Geometric	Geometric	Geometric
Sequence ratio	2	2	2

1 to the total number of variables. The R package *randomForest* (Liaw & Wiener 2002) was used to train the RF models.

Genetic programming

Genetic programming (GP) (Koza 1992) is based on an idea of randomly combining operators and elementary functions in a single formula, using the randomized search (typically using a genetic algorithm) in the space of all possible combinations and aiming at obtaining the resulting accurate formula (being a nonlinear regression equation) representing the input–output mapping. *Discipulus* software (Francone 1998) was used to implement the program-based GP. The probability of crossover and mutation was selected by combining the values 0.1, 0.3, 0.5. The final values were those minimizing the error on the cross-validation set; in this, we followed the findings of an earlier study by Elshorbagy *et al.* (2010b).

Extreme learning machines

Extreme learning machines were initially proposed as an alternative training algorithm for ANN, with the main purpose of reducing the networks to a linear system and allowing an analytical solution to the determination of the output weights (Taormina & Chau 2015b). For that to be possible, the input weights and hidden biases are randomly assigned. These simplifications drastically increase the speed of the learning process. The number of hidden nodes is the parameter to be optimized. The number of hidden neurons of each member of the ensemble was optimized from a set of values ranging from 3 to 15. The R package *elmNN* (Gosso 2012) was used to train the ELM models.

Metrics of model performance

Model performance: error statistics

The predictive capabilities of the modeling techniques used are evaluated using the root mean squared error (RMSE) and the Nash–Sutcliffe efficiency index (NSE). These comparative estimators provide information about the model's accuracy (based on the RMSE), and the model's accuracy divided by the standard deviation of the process (based on the NSE). The RMSE is, therefore, a measurement of the error variance, while the NSE provides a score in the interval $(-\infty; 1]$ for the error variance. NSE values equal to one represent a perfect predictor, while NSE values equal to zero represent the predicting capability of the average of the population. Their mathematical formulations are expressed in Equations (1) and (2):

$$RMSE = \sqrt{\frac{\sum_{i=i}^{N} (O_i - P_i)^2}{N}} \tag{1}$$

$$NSE = 1 - \frac{\sum_{i=1}^{N} (O_i - P_i)^2}{\sum_{i=1}^{N} \left(\bar{O} - O_i\right)^2} \tag{2}$$

where N is the number of instances in the set, and P_i, O_i, $\bar{P}$, and $\bar{O}$ are correspondingly the predicted variable, the observed one and their respective mean values.

Baseline: autoregressive and naïve model

Apart from the five models presented above, two simple models have also been implemented: the naïve model and the autoregressive one. These models are widely used in DDM studies as baseline models to which other models are compared. The naïve model assumes no variation in the state of the system between the subsequent time steps. In other words, given a prediction horizon i, the naïve model states that the GW depth remains the same between time t and time $t+i$. The autoregressive model is a linear estimator that assumes the output of a system y is linearly dependent on a set of previous model outputs:

$$y_t = \alpha_1 y_{t-1} + \alpha_2 y_{t-2} + \ldots + \alpha_n y_{t-n} \tag{3}$$

where n represents the order of the model (i.e., the number of lags to be considered) and $\alpha_{1\ldots n}$ are the model's coefficients, estimated by minimizing the least squares of model residuals.

In this study, the order of the AR model is progressively increased until the variation in the RMSE in the

cross-validation set becomes marginal. This procedure has led to the selection of an autoregressive model of order 2 (AR_2) for all the lead times analyzed.

Model evaluations for rising and FLs

Figure 4 shows the water table level for a yearly cycle, and it is possible to distinguish the two separate water level regimes. The first one is represented by a FL, driven by farmers' water withdrawals and a high evapotranspiration rate during the growing season. The second one, a rising limb, is driven by natural aquifer recharge, recharge from precipitation, and snowmelt. The built models are assessed for these different hydrological regimes.

Model evaluations for various water level ranges

The models are tested also for three ranges of water levels. The test set is divided into three blocks based on the 15% and the 85% quantiles of the empirical distribution of GW levels. These values are used to determine the low (LWR, in the bottom part of Figure 4), middle (MWR, in the central part of Figure 4), and high (HWR, in the upper part of Figure 4) water level ranges. The first of them is used to assess the predicting capability of the models in water shortage conditions, while the second and the third provide insights regarding the models' performances in the average and high water-level conditions.

Receiver operating characteristic

ROC curves were used to judge the discrimination ability of forecasting methods. They are a function of sensitivity and specificity. Sensitivity (SE) is the probability of obtaining a value above a certain threshold among all the cases above the threshold itself. Specificity (SP) is the probability of a negative test result, or a value below a threshold, among all the cases below the threshold itself. The ROC curve is a graphical representation of 1-SP (false positive rate) versus SE (true positive rate), and perfect discrimination ability is identified by two perpendicular lines that intersect each other in (1, 1). By using ROC to evaluate models' performances, the most important statistic is the area under the curve (AUC). The value of AUC goes from 0.5 (no discrimination ability) to 1 (100%) discrimination ability. In this study, the AUC statistic is computed for the HWR and the LWR in order to assess the discrimination ability of the models in the two different water availability conditions.

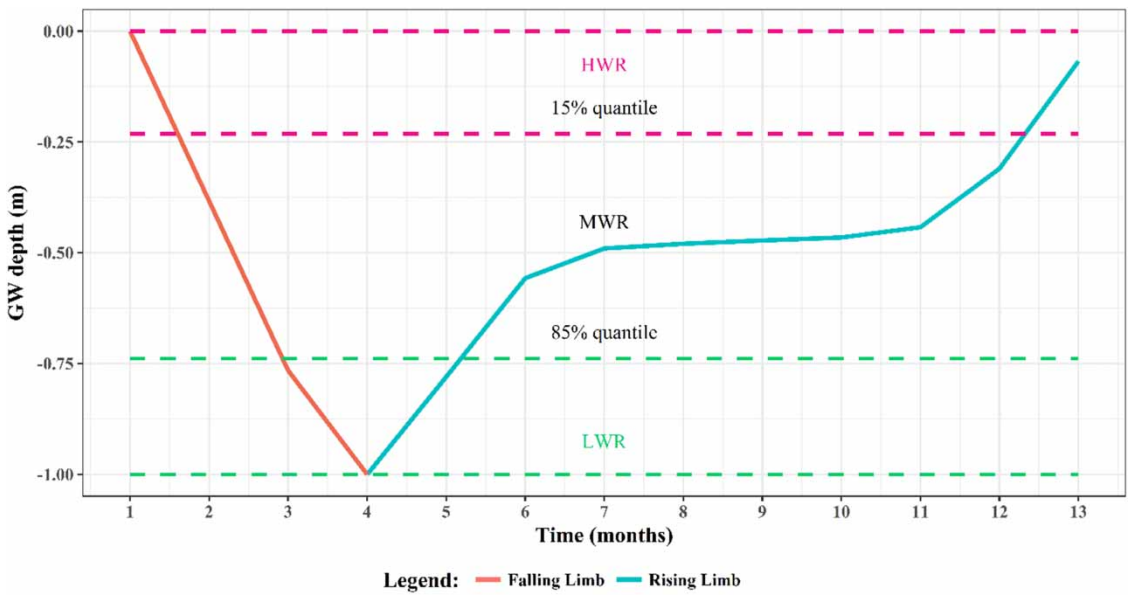

Figure 4 | Schematic representation of the FL, the rising limb and of the three water level ranges used to evaluate the different models' performances.

RESULTS AND DISCUSSION

CIVS results

Table 3 presents the variable selected by the CIVS method, for each of the lead times considered in this study. The selected variables show consistency as lead time increases. The contribution of SNM to forecast GW level changes is minimized due to the absence of snow storage in the initialization time produced by CIVS. This might be attributed to the contribution of snowmelt (occurring in March and April) to recharge as lead time reaches five months (July and August) when groundwater withdrawals are at their top, so crop irrigation requirements can be supplied. This is evidenced by the consistent sensitivity of P, ET, and AD in all lead times. Considering that pumping data are integrated to the variations in corn water demands and ET in irrigated working lands, the sensitivity analysis on this particular input results in an average increase of 15% in the NSE when AD is used. Certainly, corn water demand is fixed according to Kranz *et al.* (2008), which also constrains the ability of the model and the inherent improvement of the NSE.

To test the effectiveness of the CIVS methodology, we used the testing set to compare the NSE of the models obtained by CIVS algorithm with the one obtained by implementing the GAGRNN algorithm developed by Bowden *et al.* (2005). This algorithm uses self-organizing maps to reduce the input dimensionality and then develops a hybrid genetic algorithm and regression ANN to determine significant inputs. For the sake of the current case study, Table 3 shows that both methods were capable of capturing the nonlinearities of the system in good agreement, providing similar performances. The reduction of the performance along the forecast horizon seems not to change significantly; even LT3 and LT4 are in a very close range.

By observing the variables selected by the CIVS in Table 3, one might wonder why, for example, ET_{t+1} influences GW_{t+1} but not GW_{t+2}. Data experiments for the current case study show that including such a variable in the input set for two-month lead time will, on the one hand, ensure physical consistency, but on the other hand, leads to the selection of a least accurate model. Ensuring physical consistency or using a better model is a decision subject to the modelers' judgement. For this case study, the authors chose the most accurate model.

Models' performance without the perfect forecast

The aquifer under study being shallow, it is reasonable to suppose that interaction between future values of climatic variables and changes in GW level occurs. In the absence of weather forecasts, the authors assume perfect forecast (PF) of P, ET, and SNM. To test the validity of the assumption, a model run without using the forecasted values has also been performed, and the results were compared with those obtained with the use of forecasts (Table 4).

Analysis of the NSE values in the table shows that the use of future meteorological information marginally improves the performance of the models (0.05 average improvement). This supports the authors' opinion of fast weather–GW system interaction. However, NSE statistics for the 'no PF' model structure shows that models of good quality can still be obtained without the forecasts.

Table 3 | Variables selected by the CIVS at each lead time and NSE comparison with the GAGRNN algorithm forecasting lead time

		NSE		
τ	Output	CIVS	GAGRNN	Selected input variables (CIVS)
1	GW_{t+1}	0.93	0.93	P_{t+1}, P_t, ET_{t+1}, SNM_t, $AD_{t..t+1}$, GW_{t-1}, GW_t
2	GW_{t+2}	0.89	0.88	P_{t+1}, P_t, ET_{t-1}, SNM_t, $AD_{t..t+2}$, GW_{t-1}, GW_t
3	GW_{t+3}	0.76	0.76	P_{t+1}, P_t, ET_{t+1}, ET_t, SNM_t, $AD_{t..t+3}$, GW_{t-1}, GW_t
4	GW_{t+4}	0.73	0.70	P_{t+1}, P_t, ET_t, ET_{t-1}, SNM_{t+1}, $AD_{t..t+4}$, GW_{t-1}, GW_t
5	GW_{t+5}	0.72	0.68	P_{t+1}, P_t, ET_{t+1} $AD_{t..t+5}$, GW_{t-1}, GW_t

Table 4 | Comparison of the models results obtained with and without forecasts (NO F)

		NSE		
τ	Output	CIVS	NO F	Input variables (NO F)
1	GW_{t+1}	0.93	0.91	$P_t, SNM_t, AD_{t..t+1}, GW_{t-1}, GW_t$
2	GW_{t+2}	0.89	0.84	$P_t, ET_{t-1}, SNM_t, AD_{t..t+2}, GW_{t-1}, GW_t$
3	GW_{t+3}	0.76	0.72	$P_t, ET_t, SNM_t, AD_{t..t+3}, GW_{t-1}, GW_t$
4	GW_{t+4}	0.73	0.66	$P_t, ET_t, ET_{t-1}, AD_{t..t+4}, GW_{t-1}, GW_t$
5	GW_{t+5}	0.72	0.63	$P_t, AD_{t..t+5}, GW_{t-1}, GW_t$

Evaluation of model performance

The performance of the various DDMs for the five lead times (one to five months) of the analysis is provided in Figure 5 (here model performance is shown for the training and testing set. All subsequent figures refer to the testing set only). As expected, model performance is marginally better in the training set. Average increase of 6 cm in RMSE in the testing set supports the idea that all the models are robust against overfitting. It can be seen from Figure 5 that the performance of all models deteriorates with an increase in lead time, as is evidenced by an average decrease of 0.37 in NSE between month 1 and 5. This is particularly pronounced for the two baseline models, AR and the naïve; nonetheless, the error statistics for a lead time of one month are comparable to those of the other models (NSE baseline = 0.85; NSE DDMs = 0.88–0.94). The reason why AR and naïve models have good performance for a monthly forecast lies in the presence of an autocorrelated input (the correlation between the output and the previous month's groundwater depth is 0.9), which dominates the dynamics of the system for short lead times This is also evidenced by Elshorbagy *et al.* (2010b), who found, in a rainfall–runoff case study, that good linear models' performance occurs when systems are strongly autocorrelated. However, when the lead time increases, the effect of the autocorrelated input becomes weaker, and the performances of the two linear models with no exogenous input deteriorate much more than the others. Given the poor performances of the baseline, further analysis in this paper will be regarding only the five DDMs.

Integrating the remaining techniques into the assessment of performance, from Figure 5, we can see that none of the techniques completely outperforms the others for all the considered lead times. Differences in NSE and RMSE values hardly reach differences beyond 20%. However, it is possible to observe that EnsAnn, EnsSvm, EnsGp, and EnsElm can be considered the best predictors for this particular case in terms of accuracy with NSE about 15% above EnsRf, which provided better results than did the two linear modes, but those results still are not comparable with the ones obtained from using the other four techniques.

The ensemble averages of all the various DDMs for the five lead times of the analysis are represented in Figure 6. The performance deterioration with the increase in the forecasting horizon is clearly visible. This is particularly true in the summer of 2012, when a so-called flash drought occurred in Nebraska favoring an increase in pumping in the well assessed. In this case, while preserving good forecasting accuracy, the pumping proxy formed by integrated corn water demand, ET, and precipitation were not fully able to help in capturing the inter-seasonal variability in water abstraction, leading to an underestimation of the water table depth.

Models' performance for rising and FLs

In this study, the performance of the different models was assessed also in withdrawal and recharge-driven conditions. Withdrawal conditions are identified with the FL of the water level, usually occurring during the crop growing season. Recharge conditions reflect the rising limb, which is usually associated with autumn and winter. Figure 7 shows the performance of the models in the rising and FL of the GW table. On average, all the techniques show higher (80%) error in the FL, and lower in the rising limb. For example, one can observe that the RMSE in the rising limb goes from less than 10 cm (EnsElm) to 32 cm (EnsRf)

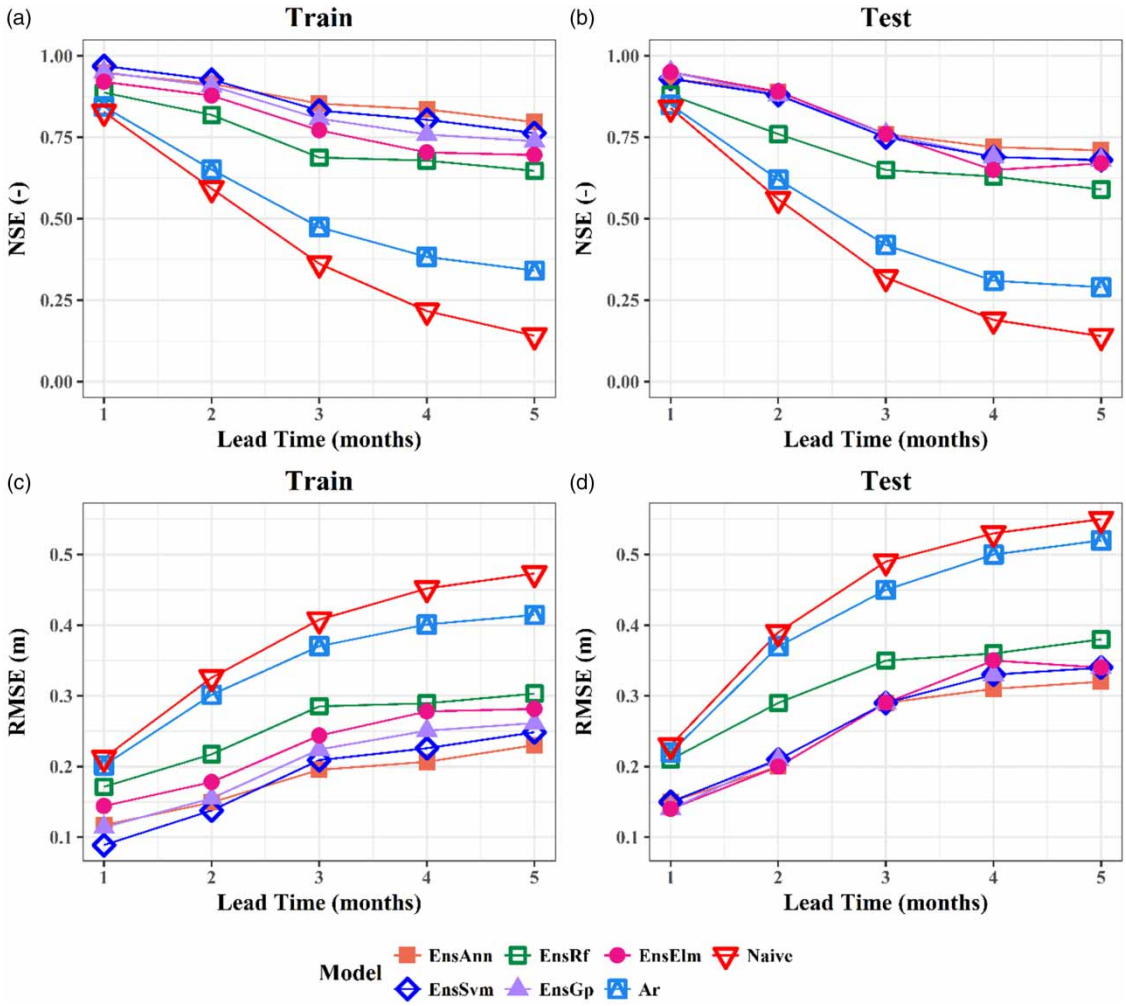

Figure 5 | Comparison of NSE (a) and (b) and RMSE (c) and (d) for the different DDMs in the training (a) and (c) and testing set (b) and (d).

for a lead time of one and five months, respectively; while in the FL for the same lead times the RMSE goes from 18 cm (EnsGp) to 45 cm (EnsRf). This phenomenon might be explained by the following.

First, the rising limb is influenced by the occurrence of natural forcings, and their influence on the GW system is easily captured by the model. On the other hand, the FL occurs under the influence of integrated natural and management-induced conditions, which are much more difficult to monitor. Please note that in the current study, the pumping data are imbedded in the integrated contributions of corn water demand, ET, and precipitation, which integrate the inter-annual variation of farmers' management practices, weather variations, and crop responses.

Second, the rising limb is recharge-driven (natural aquifer recovery, precipitation, and snowmelt), while the FL is withdrawal-driven (pumping and evapotranspiration). Usually, recharge takes place at a much lower speed than withdrawal, so the slope of the FL is usually much sharper than the rising one. This is particularly true for the winter part of the rising limb. During the winter, frost on the soil prevents precipitation and snowmelt from recharging the water table, the level of which remains almost constant.

In this comparison among the different modeling techniques, EnsRf again proved to be the worst predictor, with an RMSE value averaging 40% (rising limb) and 20% (FL) higher than the other techniques. This might be due to the

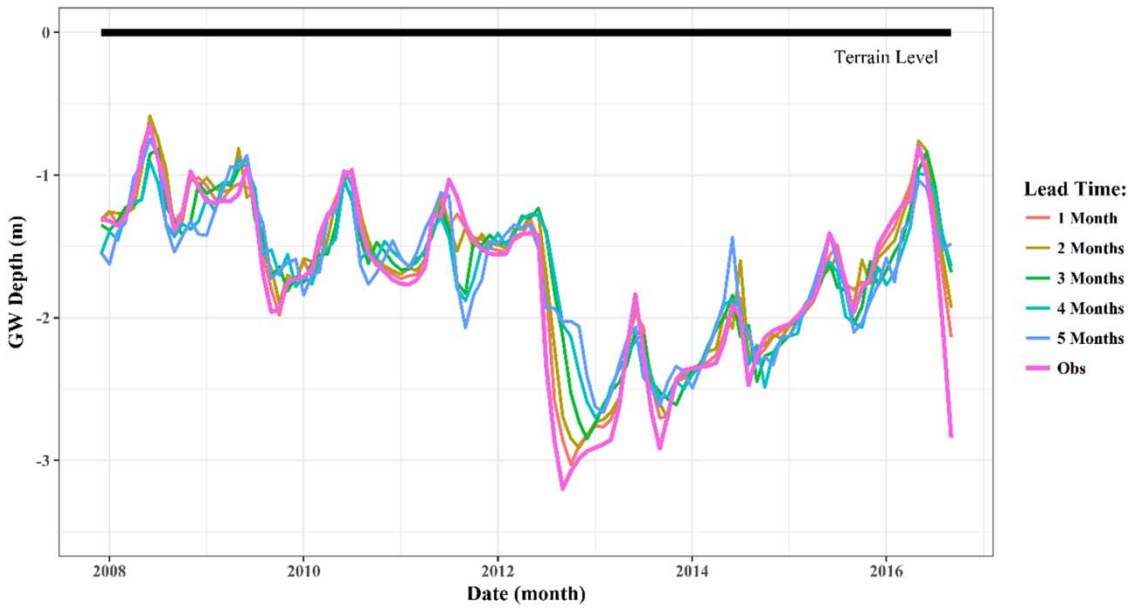

Figure 6 | Ensemble average for each lead time of the five DDMs in the testing set.

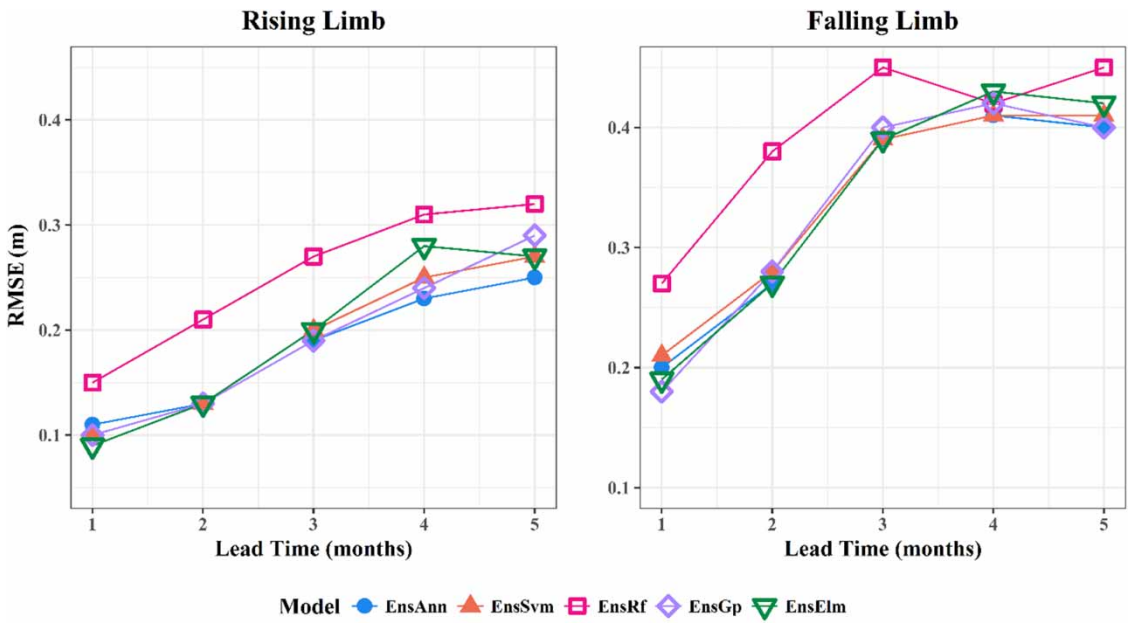

Figure 7 | RMSE in the rising and FL.

fact that EnsRf is an ensemble of several linear machines, while the other predictors are purely nonlinear. EnsAnn provides marginally better performances in the rising limb with four and five months' lead time. For any other conditions, it is impossible to establish the best predictor.

Models' performance for various water level ranges

Figure 8 shows the errors of the different techniques in three different water level ranges: the low (LWR), middle (MWR), and high (HWR) water levels. As presented earlier, they

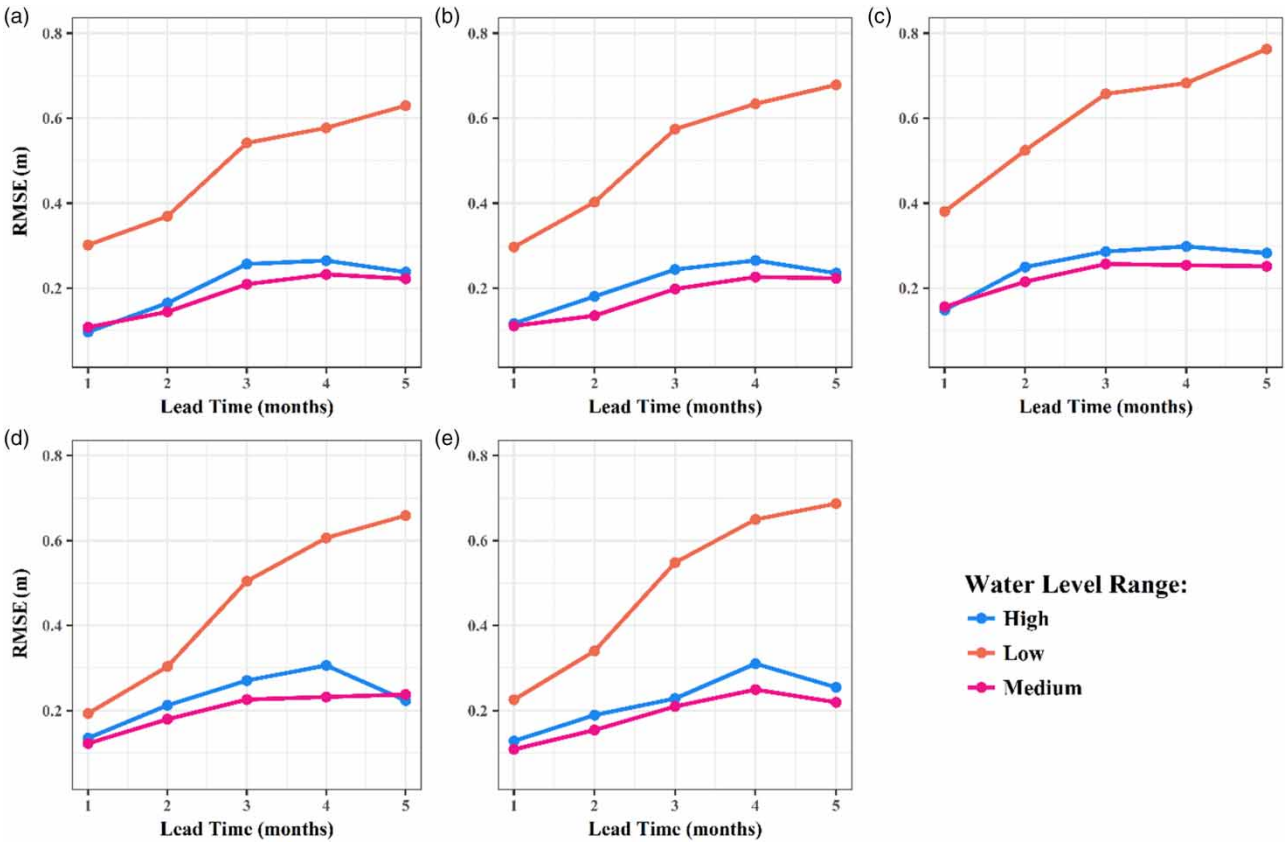

Figure 8 | RMSE plot in the HWR, MWR, and LWR. (a) EnsAnn, (b) EnsSvm, (c) EnsRf, (d) EnsGp, (e) EnsElm.

represent the GW data above the 85% quantile, intermediate, and below the 15% quantile, respectively. One may see good and similar model performances in the HWR and in the MWR. This result was expected: the MWR usually includes the winter season, when the water table rise is delayed by soil frost. The HWR occurs after the spring recharge generated by precipitation and snowmelt. Spring recharge is much faster than the winter one, and therefore, the average error increases by about 13% (from 19.5 to 22 cm). However, the error of all predictors increases in the LWR, i.e., in a water shortage situation. In particular, the RMSE range in the MWR and HWR is 9–31 cm (EnsAnn, one month's lead time; EnsRf, five months' lead time), while in the LWR it is 19–76 cm (EnsGp, one month's lead time; EnsRf, five months' lead time). The explanation lies in the dominating phenomena driving water table levels in the LWR: pumping (FL in the LWR) and natural aquifer recharge (rising limb in the LWR). They are, in fact, responsible for the fastest changes in water table level. As a consequence, when they occur, the accuracy of the model decreases. In addition, despite showing good extrapolation ability, DDMs show a decrease in accuracy when forecasting a condition not represented in the training set. In 2012, one such condition occurred; the drought of 2012 (included in the LWR), also known as flash drought was the highest recorded (NOAA 2013) since the Dust Bowl era in the 1930s (when no GW data were available).

The aforementioned findings are in good agreement with Huang *et al.* (2017). Despite the fact that a water level ranges' analysis was not carried out in their study, the analysis of the results showed an increase in the error in water shortage conditions.

Interestingly, in the HWR and in the MWR, some of the models show marginally better performance when the lead time is set to five months instead of four. Data analyses performed on the testing set lead us to believe that the explanation of this phenomenon lies in the overall 'error

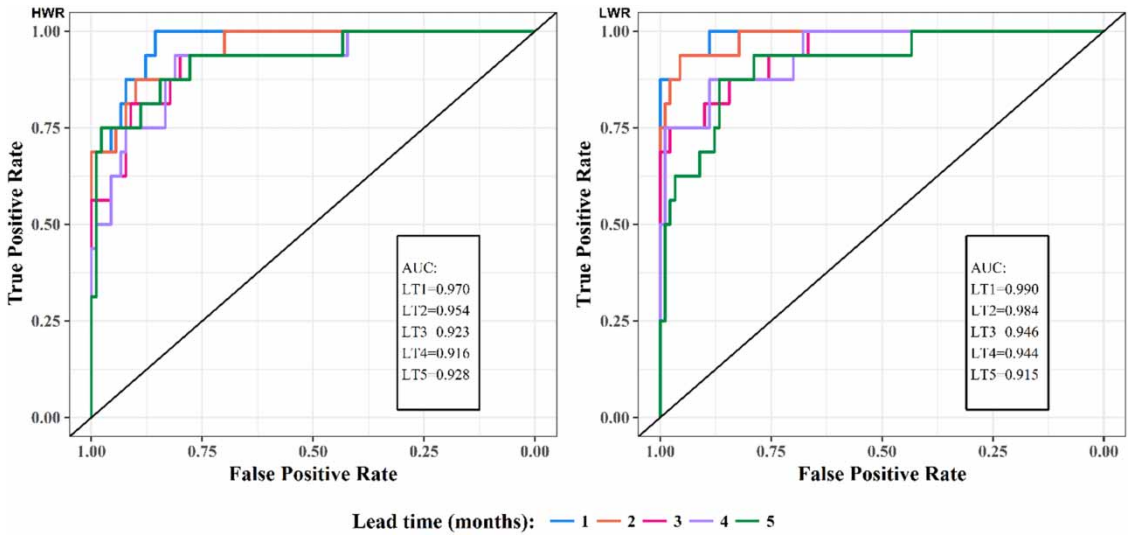

Figure 9 | EnsAnn ROC for the HWR and the LWR.

compensation' in the training process: when the lead time is four months, the higher error in the HWR and MWR is compensated by lower error in the LWR, leading to an overall better performance with respect to the five months' lead time.

Regarding the comparison among various models, EnsElm and EnsGp provide the best estimations in LWR for short lead times. The RMSE value of EnsElm and EnsGp on a monthly forecast in the LWR was about 30% lower than the one obtained with EnsAnn (RMSE = 0.19 m for EnsGp, 0.21 m for EnsElm, and 0.30 m for EnsAnn). On the other hand, EnsAnn proved the most stable technique in the LWR when the forecasting horizon increases (RMSE = 0.61 m, 0.66 m, and 0.69 m, respectively). EnsRf shows a similar performance with respect to the other predictors in the HWR and in the MWR. However, for the LWR, the RMSE value shows an average increase of 30% with respect to the other techniques. It is, therefore, very likely that the overall worst performance of EnsRf shown in Figure 5 was due to a systematic underestimation of the water table level in the LWR.

Receiver operating characteristic

Figure 9 represents the ROC curves of the five analyzed lead times in the HWR and in the LWR. The ROC of Figure 9 refers only to results obtained with EnsAnn. Similar performances were found for the other four DDMs. The AUC identifies the probability of correct classification of the aforementioned water level conditions. Analysis of the two curves shows encouraging classification accuracy (>90%), even in LWR conditions and with lead times of five months. With reference to the numerical values in Figure 9, when the lead time is increased from one to five months, the AUC value shows a decrease of 5% in the HWR and of 7% in the LWR. It is, therefore, possible to say that the deterioration in classification accuracy with increased lead time is marginal in both water level ranges. Surprisingly, as can be seen from the numerical values in Figure 9, even if the error in the LWR is usually higher, the HWR is characterized by a marginally lower classification accuracy. The explanation for this might lie in the fact that the HWR is much narrower than the LWR. In fact, the LWR includes all the water levels between −2.3 and −3.2 m below ground level (width = 0.9 m), while the HWR only includes values between −0.65 and −1.17 m below ground level (width = 0.53 m). This is also confirmed by the right-skewed distribution of the data.

CONCLUSIONS

This study assessed the ability of five data-driven techniques to forecast groundwater levels from one to five months in

different hydrogeological regimes. The predictive accuracy of the models was determined by computing three different error statistics (NSE, RMSE, and AUC) for various water level conditions. Analysis of the results, related to the hypothesis posed earlier, showed the following:

- Replacing the unknown pumping rates by a proxy (crop water demand) appeared to be useful; it increased the performance of all models.
- In this experiment, ANNs, GP, support vector machines, and extreme learning machines provided similar predicting abilities. However, ANNs performed marginally better when the lead time was increased up to five months. All the models outperformed the baseline techniques, represented by the autoregressive and the naïve models. This was particularly true with increase in the lead time and the behavior of the system becomes strongly nonlinear.
- Overall, the random forests model was the worst estimator among the five DDMs tested in this study. In particular, it systematically underestimated the water table level in conditions of water shortage.
- The RMSE of all the models was higher in the withdrawal-driven (falling) limb rather than in the recharge-driven (rising) limb. This can be explained by the high uncertainty (lack of knowledge) of the pumping patterns, which were the dominant forcing when the water table level was decreasing.
- All the models showed good agreement and encouraging performance in the high water range (HWR) and in the middle water range (MWR). The magnitude of the RMSE was higher in the low water range (LWR). In water shortage conditions, extreme learning machines and GP provided highest forecasting accuracy for short lead times, while ANNs were less sensitive to the increase in the forecasting horizon.
- Analysis of the ROC AUC statistics performed on the HWR and LWR showed a good overall ability of the models to discriminate the water level ranges under consideration, with AUC values always higher than 90%.

The results obtained from this case study are encouraging. The high error in the FL and in the LWR (when pumping takes place) can be perhaps explained by the absence of water withdrawal (pumping) data.

Despite encouraging findings, one of the limitations of the current research is the use of monthly forecasts of hydrometeorological variables as inputs. Being aware of the uncertainty associated with the estimation of those values, one of the future research efforts can be towards quantification of the sensitivity of groundwater variability to the uncertainty in the meteorological input estimation. Possible developments can be directed towards the application of the proposed methodology at the aquifer scale and the development of a guideline for the use of GW level forecasts for water management in agriculture in a wider context. While preserving good discrimination and extrapolation ability, and good forecasting accuracy, the error of the models increased during the flash drought of 2012. Considering the better capability of physically based models to extrapolate data unseen during the calibration, another possible research direction would be building composite (hybrid) models, combining data-driven and physically based approaches (distributed groundwater models), so that the best features of both would be combined. A possible architecture then could be based on the idea of the 'fuzzy committees' (Fenicia *et al.* 2007; Kayastha *et al.* 2013) which ensure a smooth shift from one model to another depending on hydrological conditions.

ACKNOWLEDGEMENTS

The authors acknowledge the support provided by the Robert B. Daugherty Water for Food Global Institute at the University of Nebraska. Some research ideas and components were also developed within the framework of the USDA National Institute of Food and Agriculture, Hatch project NEB-21-166 Accession No. 1009760 and grant No. 17-77-30006 of the Russian Science Foundation. The authors also appreciate the comments made by the reviewers and acknowledge their contribution to strength the present document.

REFERENCES

Abrahart, R. J., Anctil, F., Coulibaly, P., Dawson, C. W., Mount, N. J., See, L. M., Shamseldin, A. Y., Solomatine, D. P., Toth, E. & Wilby, R. L. 2012 Two decades of anarchy? Emerging

themes and outstanding challenges for neural network river forecasting. *Progress in Physical Geography* **36** (4), 480–513.

Akhtar, M. K., Corzo, G. A., Van Andel, S. J. & Jonoski, A. 2009 River flow forecasting with artificial neural networks using satellite observed precipitation pre-processed with flow length and travel time information: case study of the Ganges River basin. *Hydrology and Earth System Sciences* **13** (9), 1607–1618.

Bartolino, J. R. & Cunningham, W. L. 2003 *Ground-water Depletion Across the Nation*. U.S. Geological Survey Fact Sheet-103-03.

Bergmeir, C. & Benítez, J. M. 2012 Neural networks in R using the Stuttgart Neural Network Simulator: RSNNS. *Journal of Statistical Software* **46** (7), 1–26.

Bowden, G. J., Dandy, G. C. & Maier, H. R. 2005 Input determination for neural network models in water resources applications. Part 1 – background and methodology. *Journal of Hydrology* **301** (1–4), 75–92.

Breiman, L. 2001 Random forests. *Machine Learning* **45** (1), 5–32.

Campolo, M., Andreussi, P. & Soldati, A. 1999 River flood forecasting with a neural network model. *Water Resources Research* **35** (4), 1191–1197.

Coppola Jr, E., Poulton, M., Charles, E., Dustman, J. & Szidarovszky, F. 2003a Application of artificial neural networks to complex groundwater management problems. *Natural Resources Research* **12** (4), 303–320.

Coppola Jr, E., Poulton, M., Charles, E., Dustman, J. & Szidarovszky, F. 2003b Artificial neural network approach for predicting transient water levels in a multilayered groundwater system under variable state, pumping, and climate conditions. *Journal of Hydrologic Engineering* **8** (6), 348–360.

Coppola Jr, E. A., Rana, A. J., Poulton, M. M., Szidarovszky, F. & Uhl, V. W. 2005 A neural network model for predicting aquifer water level elevations. *Ground Water* **43** (2), 231–241.

Corzo, G. & Solomatine, D. 2007 Baseflow separation techniques for modular artificial neural network modelling in flow forecasting. *Hydrological Sciences Journal* **52** (3), 491–507.

Coulibaly, P., Anctil, F., Aravena, R. & Bobée, B. 2001 Artificial neural network modeling of water table depth fluctuations. *Water Resources Research* **37** (4), 885–896.

Daliakopoulos, I. N., Coulibaly, P. & Tsanis, I. K. 2005 Groundwater level forecasting using artificial neural networks. *Journal of Hydrology* **309** (1–4), 229–240.

Dibike, Y. B. & Solomatine, D. P. 2000 River flow forecasting using artificial neural networks. *Physics and Chemistry of the Earth, Part B: Hydrology, Oceans and Atmosphere* **26** (1), 1–7.

Dimitriadou, E., Hornik, K., Leisch, F., Meyer, D., Weingessel, A. & Leisch, M. F. 2009 *Package 'e1071'*. R software package. Available at: https://cran.r-project.org/package=e1071.

Djurovic, N., Domazet, M., Stricevic, R., Pocuca, V., Spalevic, V., Pivic, R., Gregoric, E. & Domazet, U. 2015 Comparison of groundwater level models based on artificial neural networks and ANFIS. *The Scientific World Journal*. Article ID 742138, 1–13.

Elshorbagy, A., Corzo, G., Srinivasulu, S. & Solomatine, D. P. 2010a Experimental investigation of the predictive capabilities of data driven modeling techniques in hydrology – Part 1: concepts and methodology. *Hydrology and Earth System Sciences* **14** (10), 1931–1941.

Elshorbagy, A., Corzo, G., Srinivasulu, S. & Solomatine, D. 2010b Experimental investigation of the predictive capabilities of data driven modeling techniques in hydrology – Part 2: application. *Hydrology and Earth System Sciences* **14** (10), 1943–1961.

Eschner, T. R., Hadley, R. F. & Crowley, K. D. 1983 *Hydrologic and Geomorphic Studies of the Platte River Basin*. U.S. Geological Survey Professional Paper 1277.

Fenicia, F., Solomatine, D. P., Savenije, H. H. G. & Matgen, P. 2007 Soft combination of local models in a multi-objective framework. *Hydrology and Earth System Sciences* **11** (6), 1797–1809.

Flowerday, C., Kuzelka, R. & Pederson, D. 1998 *The Groundwater Atlas of Nebraska*. Conservation and Survey Division, Institute of Agriculture and Natural Resources, University of Nebraska-Lincoln, Lincoln, NE.

Francone, F. D. 1998 *Discipulus Owner's Manual*. Machine Learning Technologies, Inc., Littleton, CO.

Galelli, S. & Castelletti, A. 2013 Tree-based iterative input variable selection for hydrological modeling. *Water Resources Research* **49** (7), 4295–4310.

Galelli, S., Humphrey, G. B., Maier, H. R., Castelletti, A., Dandy, G. C. & Gibbs, M. S. 2014 An evaluation framework for input variable selection algorithms for environmental data-driven models. *Environmental Modelling & Software* **62**, 33–51.

Gosso, A. 2012 *elmNN: Implementation of ELM (Extreme Learning Machine) algorithm for SLFN (Single Hidden Layer Feedforward Neural Networks)*. R package version 1.0.

Gutentag, E. D., Heimes, F. J., Krothe, N. C., Luckey, R. R. & Weeks, J. B. 1984 *Geohydrology of the High Plains Aquifer in Parts of Colorado, Kansas, Nebraska, New Mexico, Oklahoma, South Dakota, Texas, and Wyoming*. U.S. Geological Survey Professional Paper 1400.

Hanson, R. T., Schmid, W., Faunt, C. C. & Lockwood, B. 2010 Simulation and analysis of conjunctive use with MODFLOW's farm process. *Ground Water* **48** (5), 674–689.

Haykin, S. 1999 *Neural Networks: A Comprehensive Foundation*, 2nd edn. Prentice Hall, Upper Saddle River, NJ.

Huang, F., Huang, J., Jiang, S.-H. & Zhou, C. 2017 Prediction of groundwater levels using evidence of chaos and support vector machine. *Journal of Hydroinformatics* **19** (4), 586–606.

Kayastha, N., Ye, J., Fenicia, F., Kuzmin, V. & Solomatine, D. P. 2013 Fuzzy committees of specialized rainfall-runoff models: further enhancements and tests. *Hydrology and Earth System Sciences* **17** (11), 4441–4451.

Kim, T.-W. & Valdés, J. B. 2003 Nonlinear model for drought forecasting based on a conjunction of wavelet transforms and neural networks. *Journal of Hydrologic Engineering* **8** (6), 319–328.

Koza, J. R. 1992 *Genetic programming: On the Programming of Computers by Means of Natural Selection*, Vol. 1. MIT Press, Cambridge, MA.

Kranz, W. L., Irmak, S., van Donk, S. J., Yonts, C. D. & Martin, D. L. 2008 *Irrigation Management for Corn. Neb Guide*. University of Nebraska, Lincoln.

Le, M. H., Perez, G. C., Solomatine, D. & Nguyen, L. B. 2016 Meteorological drought forecasting based on climate signals using artificial neural network – a case study in Khanhhoa Province Vietnam. *Procedia Engineering* **154**, 1169–1175.

Liaw, A. & Wiener, M. 2002 Classification and regression by randomForest. *R News* **2** (3), 18–22.

Lohani, A. K. & Krishan, G. 2015 Groundwater level simulation using artificial neural network in southeast Punjab, India. *Journal of Geology and Geosciences* **4** (3), 206.

Maier, H. R. & Dandy, G. C. 2000 Neural networks for the prediction and forecasting of water resources variables: a review of modelling issues and applications. *Environmental Modelling & Software* **15** (1), 101–124.

May, R. J., Maier, H. R., Dandy, G. C. & Fernando, T. G. 2008 Non-linear variable selection for artificial neural networks using partial mutual information. *Environmental Modelling & Software* **23** (10–11), 1312–1326.

Mohanty, S., Jha, M. K., Kumar, A. & Sudheer, K. P. 2010 Artificial neural network modeling for groundwater level forecasting in a river island of eastern India. *Water Resources Management* **24** (9), 1845–1865.

Mohanty, S., Jha, M. K., Raul, S. K., Panda, R. K. & Sudheer, K. P. 2015 Using artificial neural network approach for simultaneous forecasting of weekly groundwater levels at multiple sites. *Water Resources Management* **29** (15), 5521–5532.

NAS-USDA 2011 *USDA's National Agricultural Statistics Service, Northern Plains Regional Field Office*. https://www.nass.usda.gov/Statistics_by_State/Nebraska/index.php (accessed 10 September 2017).

Nayak, P. C., Rao, Y. S. & Sudheer, K. 2006 Groundwater level forecasting in a shallow aquifer using artificial neural network approach. *Water Resources Management* **20** (1), 77–90.

NOAA National Centers for Environmental Information 2013 *State of the Climate: Drought – Annual 2012*. https://www.ncdc.noaa.gov/sotc/drought/201213 (accessed 10 September 2017).

Rodell, M., Houser, P., Jambor, U., Gottschalck, J., Mitchell, C.-J., Meng, K., Arsenault, K., Cosgrove, B., Radakovich, J., Bosilovich, M., Entin, J. K., Walker, J. P., Lohmann, D. & Toll, D. 2004 The global land data assimilation system. *Bulletin of the American Meteorological Society* **85** (3), 381.

Scanlon, B. R., Faunt, C. C., Longuevergne, L., Reedy, R. C., Alley, W. M., McGuire, V. L. & McMahon, P. B. 2012 Groundwater depletion and sustainability of irrigation in the US High Plains and Central Valley. *Proceedings of the National Academy of Sciences* **109** (24), 9320–9325.

Solomatine, D. P. & Dulal, K. N. 2003 Model trees as an alternative to neural networks in rainfall-runoff modelling. *Hydrological Sciences Journal* **48** (3), 399–411.

Solomatine, D. P. & Xue, Y. 2004 M5 model trees and neural networks: application to flood forecasting in the upper reach of the Huai River in China. *Journal of Hydrologic Engineering* **9** (6), 491–501.

Solomatine, D. P., See, L. M. & Abrahart, R. J. 2009 *Data-driven Modelling: Concepts, Approaches and Experiences*. *Practical Hydroinformatics*, Water Science and Technology Library, Vol. 68. Springer, Berlin, Heidelberg, pp. 17–30.

Sun, Y., Wendi, D., Kim, D. E. & Liong, S.-Y. 2016 Technical note: application of artificial neural networks in groundwater table forecasting – a case study in a Singapore swamp forest. *Hydrology and Earth System Sciences* **20** (4), 1405–1412.

Taormina, R. & Chau, K.-W. 2015a Neural network river forecasting with multi-objective fully informed particle swarm optimization. *Journal of Hydroinformatics* **17** (1), 99–113.

Taormina, R. & Chau, K.-W. 2015b Data-driven input variable selection for rainfall–runoff modeling using binary-coded particle swarm optimization and Extreme Learning Machines. *Journal of Hydrology* **529**, 1617–1632.

Tsanis, I. K., Coulibaly, P. & Daliakopoulos, I. N. 2008 Improving groundwater level forecasting with a feedforward neural network and linearly regressed projected precipitation. *Journal of Hydroinformatics* **10** (4), 317–330.

USGS 2015 *National Water Information System: USGS Groundwater Data for the Nation*. https://waterdata.usgs.gov/nwis/gw (accessed 4 December 2017).

Vapnik, V. 1998 *Statistical Learning Theory*, Vol. 1. Wiley, New York.

Vapnik, V. 2013 *The Nature of Statistical Learning Theory*. Springer Science & Business Media, New York.

Varouchakis, E. A. 2016 Modeling of temporal groundwater level variations based on a Kalman filter adaptation algorithm with exogenous inputs. *Journal of Hydroinformatics* **19** (2), 191–206.

Wen, F. & Chen, X. 2006 Evaluation of the impact of groundwater irrigation on streamflow in Nebraska. *Journal of Hydrology* **327** (3), 603–617.

Witten, I. H. & Frank, E. 2005 *Data Mining: Practical Machine Learning Tools and Techniques*. Morgan Kaufmann Publishers, San Francisco, CA.

First received 18 December 2017; accepted in revised form 2 May 2018. Available online 17 May 2018