A Spatially Enhanced Data-Driven Multimodel to Improve Semiseasonal Groundwater Forecasts in the High Plains Aquifer, USA

Amaranto, A.; Munoz-Arriola, F.; Solomatine, D. P.; Corzo, G.

DOI
10.1029/2018WR024301

Publication date
2019

Document Version
Final published version

Published in
Water Resources Research

Citation (APA)
Amaranto, A., Munoz-Arriola, F., Solomatine, D. P., & Corzo, G. (2019). A Spatially Enhanced Data-Driven Multimodel to Improve Semiseasonal Groundwater Forecasts in the High Plains Aquifer, USA. *Water Resources Research*, *55*(7), 5941-5961. https://doi.org/10.1029/2018WR024301

Water Resources Research

RESEARCH ARTICLE
10.1029/2018WR024301

Special Section:
Big Data & Machine Learning in Water Sciences: Recent Progress and Their Use in Advancing Science

A Spatially Enhanced Data-Driven Multimodel to Improve Semiseasonal Groundwater Forecasts in the High Plains Aquifer, USA

A. Amaranto[1,2], F. Munoz-Arriola[1,3], D. P. Solomatine[2,4,5], and G. Corzo[2]

[1]Department of Biological Systems Engineering, University of Nebraska-Lincoln, Lincoln, NE, USA, [2]Hydroinformatics Chair Group, IHE-Delft, Institute for Water Education, Delft, The Netherlands, [3]School of Natural Resources, University of Nebraska-Lincoln, Lincoln, NE, USA, [4]Water Resources Section, Delft University of Technology, Delft, The Netherlands, [5]Flood Hydrology Lab, Water Problems Institute of RAS, Moscow, Russia

Key Points:
- Artificial neural networks can accurately forecast semiseasonal groundwater level changes
- MuMoC improved the groundwater well-level forecasting skill for 1- to 4-month lead times with respect to a single ANN model by 25% in NSE
- The implementation of MuMoC is recommended in case of densely gauged areas

Correspondence to:
F. Munoz-Arriola,
fmunoz@unl.edu

Citation:
Amaranto, A., Munoz-Arriola, F., Solomatine, D. P., & Corzo, G. (2019). A spatially enhanced data-driven multimodel to improve semiseasonal groundwater forecasts in the High Plains aquifer, USA. *Water Resources Research*, *55*, 5941–5961. https://doi.org/10.1029/2018WR024301

Received 24 OCT 2018
Accepted 6 JUN 2019
Accepted article online 20 JUN 2019
Published online 24 JUL 2019

Abstract The aim of this paper is to improve semiseasonal forecast of groundwater availability in response to climate variables, surface water availability, groundwater level variations, and human water management using a two-step data-driven modeling approach. First, we implement an ensemble of artificial neural networks (ANNs) for the 300 wells across the High Plains aquifer (USA). The modeling framework includes a method to choose the most relevant input variables and time lags; an assessment of the effect of exogenous variables on the predictive capabilities of models; and the estimation of the forecast skill based on the Nash-Sutcliffe efficiency (NSE) index, the normalized root mean square error, and the coefficient of determination (R^2). Then, for the ANNs with low- accuracy, a MultiModel Combination (MuMoC) based on a hybrid of ANN and an instance-based learning method is applied. MuMoC uses forecasts from neighboring wells to improve the accuracy of ANNs. An exhaustive-search optimization algorithm is employed to select the best neighboring wells based on the cross correlation and predictive accuracy criteria. The results show high average ANN forecasting skills across the aquifer (average NSE > 0.9). Spatially distributed metrics of performance showed also higher error in areas of strong interaction between hydrometeorological forcings, irrigation intensity, and the aquifer. In those areas, the integration of the spatial information into MuMoC leads to an improvement of the model accuracy (NSE increased by 0.12), with peaks higher than 0.3 when the optimization objectives for selecting the neighbors were maximized.tT

1. Introduction

Growing demands for agricultural water increase the stress on supplies as the population grows and the climate becomes more volatile (Iglesias & Garrote, 2015; Portmann et al., 2010). In this variable supply-and-demand trade-off, groundwater (GW) helps sustain a consistent intensification of agricultural productivity around the world. However, GW withdrawals have also led to a GW depletion of 283 km^3/year worldwide (Wada et al., 2010). Consequences of aquifer overexploitation span from drying up of wells, reduction of water in streams and lakes, and water quality degradation to increased pumping costs, land subsidence, decreased well yields, and water rationing (Bartolino & Cunningham, 2003; Nayak et al., 2006). Dalin et al. (2018) and Butler et al. (2018) evidence irrigation as one of the main drivers of agriculture's sustainability and GW depletion. The effective management of water resources is an imperative task (Galelli et al., 2010) to be approached with various time scales in mind. In particular, water resources reallocations are planned semiseasonally to seasonally to optimize water use efficiency and maintain soil field capacity in the agricultural working lands and sustain water systems.

Water management encompasses social (di Baldassarre et al., 2013), economical (Giuliani et al., 2014), and operational (Giuliani et al., 2015) aspects. From an operational perspective, water table forecasts are fundamental to implementing optimal GW management policies and to conserving water resources (Coppola et al., 2005) across geopolitical and geophysical limits. With the aim of providing accurate forecasts, in the past two decades the use of data-driven models (DDMs) in the hydrological field has expanded (for a review, see Abrahart et al., 2012), with studies on rainfall-runoff modeling (Solomatine & Dulal, 2003), flood (Solomatine & Xue, 2004), and drought forecasting (Le et al., 2016). There are also DDM applications in GW, for example, by Coppola et al. (2003), who studied the ability of artificial neural networks (ANNs) to predict water table levels with a lead time (LT) of 30 days near a public supply wellfield. Sun (2013) applied ANNs to predict GW level changes using GRACE and PRISM data as inputs across the US. The use of spatially distributed inputs also evidenced the potential of coupling data driven models with spatial interpolation techniques. Tapoglou et al. (2014) implemented a hybrid ANN-Kriging model to simulate daily GW level variations across the Isar River in Bavaria, Germany (7,100 km^2). They concluded that the ANN-Kriging approach could be successfully used in aquifers where the hydrogeological information is constrained. Sun et al. (2016) analyzed the ability of ANNs to predict water table depth in a swamp forest in Singapore, establishing that accurate estimates could be obtained with a daily LT, whereas the performance decreased for the LT of a week. Yadav et al. (2017) compared the performance of extreme learning machines and support vector machines in forecasting monthly GW levels in two different wells in Canada, discovering that extreme learning machines outperformed support vector machines in both analyzed case studies. We carried out a study to compare the predicting capabilities of five different DDMs to forecast seasonal (1- to 4-month) GW levels in different hydrological regimes (Amaranto et al., 2018). It was found that all the DDMs outperformed baseline models (autoregressive and naïve) and that the error increased in water deficit conditions. Sahoo et al. (2017) used different machine learning (ML) algorithms to predict water level changes in the High Plains (HP) aquifer and the Missouri River Basin (USA), establishing that the best results are obtained when decomposing the input using spectrum analysis before forcing the ANN. The authors also concluded that ANNs outperformed hybrid linear and nonlinear regression models. Guzman et al. (2017) implemented nonlinear autoregressive neural networks (NARX) to forecast the daily GW level in a well in the Mississippi River Valley aquifer. Wunsch et al. (2018) used NARX for monthly (1 to 6 months) GW level forecasts in several wells in southwest Germany. Both obtained encouraging results, indicating the suitability of NARX for predicting GW levels. Rakhshandehroo et al. (2018) used wavelet ANNs to predict the GW level in a shallow well in Florida and a deep well in Arkansas, concluding that noisy GW fluctuations in the shallow well caused higher error, which led to their obtaining the highest accuracy for the deep well.

Despite encouraging results obtained at the basin and hydrogeological unit scales, few applications in scientific literature address implementing GW forecasting systems at the (large) aquifer scale (see, e.g., Sahoo et al., 2017). At the large scale, different land use and hydrometeorological conditions might occur; and their relationship with the modeling forecasting skills (i.e., how the predicting accuracy changes at the occurrence of such different conditions) at different LTs is yet to be understood and quantified.

Furthermore, to the best of the authors' knowledge, little has been done in designing and testing alternative data-driven modeling strategies aimed at improving the traditionally used models, which would aggregate, for example, regional geospatial GW information into a model. Our experience and results of other studies show that in many cases there might be a poor (temporal) autocorrelation in the GW-related time series, which leads to inaccuracies in ANN models built for a single location. In this situation, the inclusion into the model of spatially distributed information (including information from other models) might improve its forecasting skills, especially in case of high spatial correlation between GW signals at various locations.

The objectives of this research are as follows:

1. to analyse the accuracy of existing models (ANN) at different locations and LTs, taking into account heterogeneity in land use weather and hydrogeological conditions occurring through the aquifer;
2. to explore a hybrid multimodeling approach, combining ANN models and instance-based learning (IBL) techniques, combining forecasts from several (optimally) selected neighboring wells, and thus overcoming the limitations of single models (this multimodel will be referred to as a MultiModel Combination, or MuMoC); and
3. to develop a GW forecasting framework to predict semi-seasonal (1- to 4-month) water level changes, at the large aquifer scale.

The hypotheses associated with this study can be formulated as follows: (1) cross-space estimations of efficiency indices (Nash-Sutcliffe efficiency [NSE] index, root mean square error [RMSE], and R^2, calculated by comparing observed versus predicted data) will allow for determining the spatial distribution of semiseasonal forecasting skills of GW well-level changes in the HP aquifer; and (2) an increase in the forecast skill of a single model, represented by an improved efficiency index, will be achieved through the aggregation of the spatial and temporal inputs in MuMoC, by combining local ANN models with outputs from models for other locations, and will be proportional to the quality of these latter models.

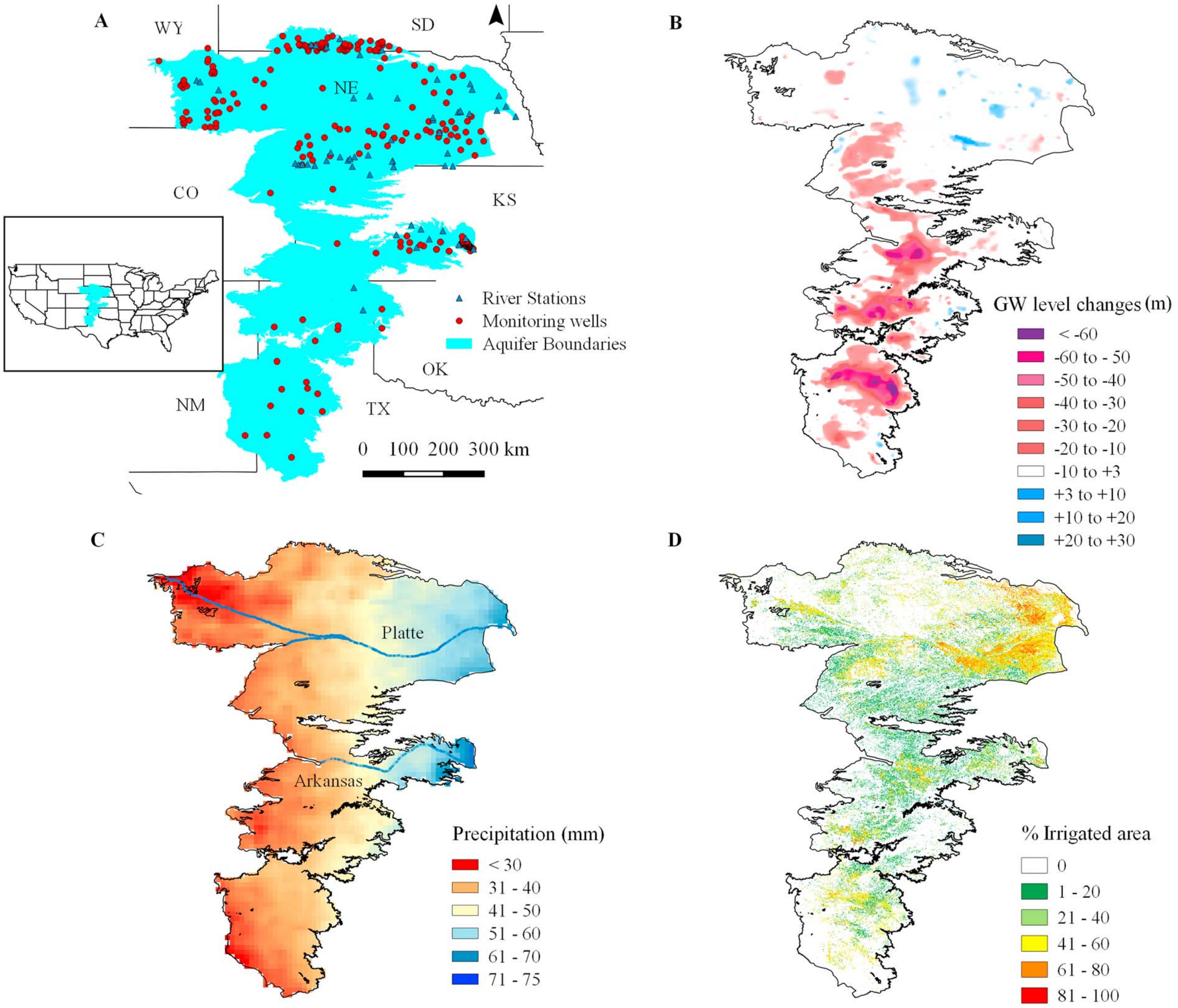

Figure 1. (a) High Plains aquifer monitoring network; (b) groundwater (GW) level changes in the 1950–2013 period (McGuire, 2017); (c) average monthly precipitation (mm) in the 1980–2016 period; and (d) percentage of irrigated area (Ozdogan & Gutman, 2008).

2. Material

2.1. Study Area and Available Data

The HP aquifer (Figure 1a) extends for 450,000 km^2 in the central part of the United States. It underlies parts of eight states: Colorado, Kansas, Nebraska, Oklahoma, South Dakota, Texas, Wyoming, and New Mexico. As can be seen from Table 1, Nebraska occupies the largest portion of the aquifer (37% of the total area), followed by Texas (20%) and Kansas (18%). The aquifer consists of hydraulically connected geologic units of later Tertiary or Quaternary age (Gutentag et al., 1984). Quaternary deposits are mainly alluvial, dune-sand,

Table 1
Area in Each of the Eight States That Belong to the HP Aquifer and Percentage of the Total HP Aquifer Area

	State							
	SD	NE	CO	KS	WY	OK	NM	TX
Area (km^2)	12,779	167,302	34,462	79,921	20,950	19,341	24,670	93,170
% total	2.8	36.9	7.6	17.6	5	4.2	5.4	20

Note. HHigh Plains; SD = South Dakota; NE = Nebraska; CO = Colorado; KS = Kansas; WY = Wyoming; OK = Oklahoma; NM = New Mexico; TX = Texas.

and valley-fill deposits. Tertiary rocks include the Brule Formation, the Arikaree group, and the Ogallala Formation. The Ogallala Formation covers about 342,000 km^2 (75% of the total aquifer area). The aquifer presents itself mainly in unconfined conditions, and its saturated thickness ranges from less than 20 m to more than 400 m in central Nebraska (McGuire, 2017).

Starting in the 1950s (a period also referred to as predevelopment), agriculture experienced a major growth, and now the area overlying the HP aquifer is one of the most developed agricultural landscapes in the United States. The National Agricultural Statistics Service (2011) estimated that the market value of the agricultural products in the HP aquifer is about $35 billion per year. According to Maupin and Barber (2005), the HP aquifer ranks first in the United States for total GW withdrawals. This caused a GW depletion in the HP aquifer of about 330 km^3 in the past 70 years, corresponding to about 8% of the total GW storage before predevelopment (McGuire, 2011). As can be seen from Figure 1b, GW depletion is not uniform through the HP aquifer; it is negligible in the north portion (Nebraska averages ~0.3 m) and much greater in the central and south portions (Kansas averages ~7 m, and Texas averages ~11 m). Scanlon et al. (2012) studied the spatial distribution of the depletion rates from 1997 to 2007 and, by extrapolating the depletion trend, they estimated that the saturated thickness of the HP aquifer could drop to less than 6 m in 35% of the southern HP aquifer within the following 30 years, rendering those areas incapable of supporting irrigation. Spatial variation in GW depletion may reflect spatiotemporal differences in water demands (WDs) by irrigation and supplies through recharge (Scanlon et al., 2012). Recharge in the HP aquifer is mainly driven by precipitation, and the surface water-GW drainage is limited to a few rivers (e.g., the Platte in Nebraska and the Arkansas in Kansas). Precipitation (Figure 1c) follows a west-to-east gradient; it is at maximum in eastern Nebraska and eastern Kansas (60–75 mm/month) and at minimum in Wyoming, Texas, and New Mexico (less than 30 mm/month). Houston et al. (2013) computed the net aquifer recharge in the years 2000–2009, estimating a maximum recharge rate of about 22 mm/year occurring in the eastern part of Nebraska and alongside the Arkansas River. A minimum recharge rate of less than 4 mm/year occurs in South Dakota, western Kansas (except the area alongside the Arkansas River), New Mexico, and Texas. Scanlon et al. (2012) report estimates of recharge rates in the HP aquifer of about 92 mm/year in the Sand Hills area (Nebraska) and a recharge rate smaller than the withdrawal rate by almost a factor of 10 in some areas of Texas and Kansas. Irrigation intensity (Figure 1d; Ozdogan & Gutman, 2008) is highest in eastern Nebraska (almost 100% of irrigated area) but is also high in Kansas and Texas, while agriculture in South Dakota is mainly rain fed.

The Global Land Data Assimilation System developed by Rodell et al. (2004) provided a monthly estimation of precipitation (P, mm/month) and evapotranspiration (ET, mm/month), with a spatial resolution of 1/8° latitude × longitude (about 15 × 15 km). Pumping data were unavailable for the HP aquifer. However, a study by Amaranto et al. (2018) found that including the crop WD as an input variable, improves ANN skills in GW level forecasting by about 20%. So the current study also uses the crop WD (mm/month) as a proxy to represent GW withdrawals. To estimate the WD for the six major crops in the HP aquifer (corn, soybeans, wheat, cotton, alfalfa, and sorghum), we applied the Food and Agriculture Organization of the United Nations (FAO) 56 methodology (Allen et al., 1998). Hence, the demand is computed as a product of reference ET and crop coefficients (which depend on crop type, sowing date, and harvesting date). Reference ET is computed with the Blaney-Criddle method (Blaney & Criddle, 1962), which takes latitude and temperature as inputs. Crop parameters were obtained from the FAO 56 database.

The U.S. Geological Survey (2015) has monitored GW (in meters below land surface) and discharge (Q, m^3/day) in the HP aquifer. In this study, we filtered the complete U.S. Geological Survey GW database in order to include stations with a minimum observation requirement of 10 years of data (120 observations) and missing data no higher than 25%. After we implemented the filter, 300 wells remained available for analysis (Figure 1a). Discharge data were extracted for the stream gauges closest to the selected monitoring wells. Summary of all variables used in the study, along with the temporal resolution, timeline, and source is presented in Table 2.

3. Methodology

To achieve the objectives described above, we designed and implemented a data-driven forecast framework (Figure 2) on 300 wells across the HP aquifer. First, we divided the collected P, ET, Q, WD, and GW level data

Table 2
Summary of Input and Output Variables: Units, Temporal Resolution, Time Span, and Data Sources

Variable	Units	Resolution	Time span	Source
GW	m	Monthly	Jan 1980/Nov 2017	https://waterdata.usgs.gov/nwis/gw
P	mm/month	monthly	Jan 1980/Nov 2017	https://ldas.gsfc.nasa.gov/nldas/NLDAS2forcing_download.php
ET	mm/month	monthly	Jan 1980/Nov 2017	https://ldas.gsfc.nasa.gov/nldas/NLDAS2model_download.php
Q	m^3/day	daily	Jan 1980/Nov 2017	https://waterdata.usgs.gov/nwis/gw
WD	mm/month	monthly	Jan 1980/Nov 2017	http://www.fao.org/docrep/X0490E/x0490e07.htm

Note. GW = groundwater depth; precipitation; ET = evapotranspiration; Q = streamflow; WD = water demand.

into training and test sets. We normalized the training and test sets between 0 and 1 with the minimum and maximum values of the former (see section 3.1). To select the most relevant input variables and lags, we applied the model-based input variable selection (IVS) procedure developed by Amaranto et al. (2018; using ANN as the model) to the training set (see section 3.2). The resulting variables from the IVS were then used to force the models.

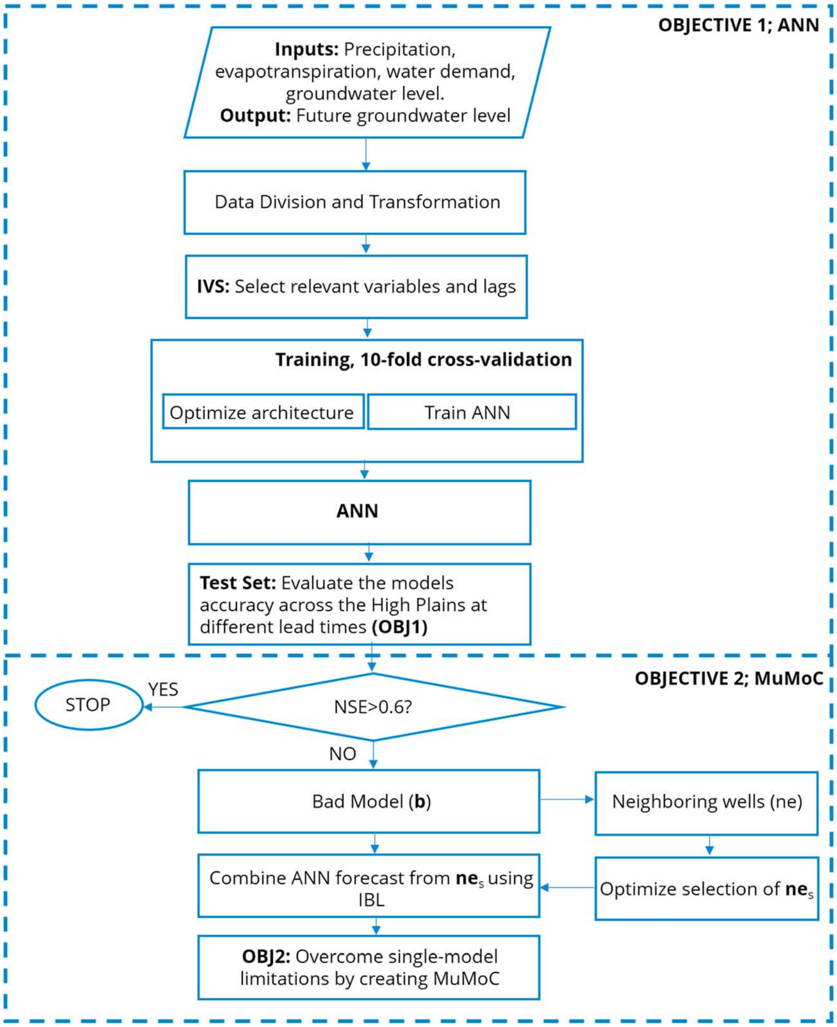
OBJECTIVE 1; ANN
Inputs: Precipitation, evapotranspiration, water demand, groundwater level.
Output: Future groundwater level
Data Division and Transformation
IVS: Select relevant variables and lags
Training, 10-fold cross-validation
Optimize architecture
Train ANN
ANN
Test Set: Evaluate the models accuracy across the High Plains at different lead times (OBJ1)
OBJECTIVE 2; MuMoC
STOP
YES
NSE>0.6?
NO
Bad Model (b)
Neighboring wells (ne)
Combine ANN forecast from ne$_s$ using IBL
Optimize selection of ne$_s$
OBJ2: Overcome single-model limitations by creating MuMoC

Figure 2. Methodological framework. (top) ANN, applied to all 300 wells in the area. (bottom) Applied for all the *b* wells. Sections 3.1 to 3.4 provide a detailed explanation of each of the blocks in the flowchart. ANN = artificial neural network; IBL = instance-based learning; IVS = input variable selection; MuMoC = MultiModel Combination; NSE = Nash-Sutcliffe efficiency.

To optimize the architecture (number of neurons) of the ANN (se section 3.3), we further divided the training set into training proper and cross-validation sets. The RMSE on the cross-validation set was used to define the number of nodes in the hidden layer. This procedure is called 10-fold cross-validation because the training is repeated 10 times for various splits (in the proportion 9:1) of the training data set, every time the ANN architecture is optimized and an ensemble of 10 predictors is generated (one for each fold). Finally, we use the test set to evaluate the performance of the forecasts (NSE) of the ensemble average.

If the performance on the test set is unsatisfactory (NSE $<$ 0.6), the model is classified as "bad" (b), and MuMoC, a hybrid ANN-IBL model (see section 3.4), is implemented. Even though there are no standard criteria for assessing a model's performance, for typical hydrological modeling, Moriasi et al. (2007), Christiansen (2012), and Me et al. (2015) proposed values of the NSE index above 0.5. To be conservative, this study raised the proposed threshold up to 0.6. The main purpose of MuMoC is to combine the GW forecast from "neighboring" wells (from hereon referred to as ne) to produce a more accurate forecast in b. To do so, we developed an exhaustive-search optimization algorithm to select the best wells (ne_{s}) geolocalized near b (see section 3.4.1). The criteria used for the selection of a certain well are the forecast skill of the model in ne wells and the cross correlation between the GW level trajectories in ne and b. Then, ANN forecasts from ne_s are combined by using IBL to produce GW forecasts in b (see section 3.4.2). The effectiveness of the MuMoC method is assessed by comparing the extent of the improvement in the resultant NSE values with those obtained with the single ANN.

3.1. Data Division and Normalization

The implementation of a DDM requires the output in the training and the test sets to have similar statistical distributions (Bhattacharya & Solomatine, 2006). This usually entails performing several random splits of data and then comparing the statistical properties (e.g., mean and standard deviation) of the training and testing sets in each split. In this study, each data set (one per well) was randomly sampled 100 times, thus creating 100 data set realizations of the output value. At each split, the mean and standard deviation of the resulting splits were computed, as well as the ratios μ_R and σ_R between the normalized means and standard deviations of the training set testing set correspondingly. Ideally, the training and test sets would have the similar mean and standard deviation, so μ_R and σ_R would be close to 1. Therefore, the split that minimizes the Euclidean distance (defined as $dis = \sqrt[2]{(\mu_R{-}1)^2 + (\sigma_R{-}1)^2}$) was used to divide the data.

The iterative random split described above minimizes the Euclidean distance between training and testing average and standard deviation. Therefore, it increases the likelihood of the models to be trained on hydrological conditions that are similar to those occurring in validation. However, in the hydrological field, it is often required to test the predictor on the most recent data. Consequently, we added a second experiment where the initial 70% of the data were used for training the model, and the remaining 30% were used to test them.

3.2. Input Variable Selection

A key step in building DDMs is selecting relevant input variables and time lags, a procedure commonly known as IVS. To do this, one could perform an exhaustive search on the input space. However, when the number of candidate input variables (and time lags) is high, an exhaustive search can be computationally expensive: If n is the number of candidate inputs, there are $2^n - 1$ possible combinations. When considering also the appropriate lags to be chosen, the complexity of the problem increases further to $(2^n - 1)^{\mathrm{lags}}$, and this makes pure exhaustive search an option only for problems with a small number of inputs. Use of (nonexhaustive) optimization approaches allows for a much more efficient search. For example, Bowden et al. (2005) proposed to use genetic algorithm for this purpose (they termed it GAGRNN, since it was applied to a general regression neural network). Elshorbagy et al. (2010a) suggested using partial mutual information and cross correlation as criteria for selection, and Galelli and Castelletti (2013) applied 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 an adaptation of the constrained input variable selection (CIVS), a methodology developed and implemented by Amaranto et al. (2018). CIVS is an exhaustive search that is however constrained by rules based on knowledge of GW physics. Consequently, instead of testing every possible input combination, models' performance is evaluated only for those combinations that respects a set of predefined rules. The original CIVS rules were developed for a single-well case study in the HP aquifer, in a corn-irrigated area where the aquifer was unconfined and particularly shallow. Here CIVS is expanded to 300 wells across the HP aquifer characterized by strong spatial heterogeneity in weather, land use, and hydrogeological properties. The rules below (more general with respect to the original CIVS rules) were implemented on P, ET, WD, Q, and GW in each of the 300 wells for the analyzed 1-, 2-, 3- and 4-month LTs.

1. The number of time lags for the autoregressive GW term is equal to the order of the autoregressive model, after which, with an increase in its order, there is no improvement in the RMSE on the cross-validation set.

Deciding the number of autoregressive terms for a DDM is a complex procedure for which no clear rules have been defined. For example, Solomatine et al. (2008) used cross correlation as a criterion for streamflow forecast applications in two different rivers, selecting only the autoregressive terms having input-output correlations above 0.9 and 0.8 (the last two lags), respectively. Shiri and KişI (2011) run a genetic programming model testing up to four autoregressive terms. In this present study, we set the number of autoregressive terms based on modeling results.

2. The maximum number of time lags for WD and ET is equal to 3, and the only lagged variables included are x_t (where x is either WD or ET), x_{t-1}, and x_{t-2} (as described in Amaranto et al., 2018).

WD and ET are variables that represent the evapotranspirative requirements, which in irrigated landscapes are proxies of unavailable GW pumping data. GW well level changes in response to GW withdrawals for irrigation in unconfined aquifers can be noted within a few days to 1 month. McMahon et al. (2011) support such assumptions in their analyses on GW recharge. Since some of the wells in the current analysis were located in areas with high water table depth, the time in this study was extended up to 3 months.

3. The maximum number of lags for P is equal to 12. However, only three of the 12 lags are tested in the exhaustive search. The three tested lags are those corresponding to the 3-month P with higher cross correlation with the outputs.

Given the high spatial heterogeneity in soil properties and water table depths, we can assume that precipitation-induced recharge occurs at different rates in different locations. Therefore, information on P from the previous year was included as an input candidate. However, to limit the search, only the 3-month P carrying the maximum amount of information (i.e., maximum cross correlation with the output) was tested. In other words, for each well, the cross correlation between each rainfall and GW level was computed. The rainfall trimester having the maximum cross correlation with the output was selected and tested in the constrained exhaustive search.

4. At least two exogenous variables among P, ET, and WD must be considered in the input set at the same time.

The reasons behind the implementation of this rule are twofold: (a) The rule excludes from all the combinations to test all the input candidates of size 2 (i.e., being $x_{t-\tau}$ any of the aforementioned variables at any considered lag τ, all the combinations including only [GW and $x_{t-\tau}$] are not considered). Therefore, it reduces computational time. (b) The HP is an aquifer heavily used for irrigation, whose dynamics are governed by the interaction of natural (P and ET) and human intervention-related (WD, and also ET if we consider that the ET demands depend on agriculture) variables. Consequently, this rule excludes all the candidates not including both natural and anthropogenic factors.

5. Lag “jumps” (or gaps) are not allowed. This means that if x_t is considered as an input, then x_{t-2} cannot be an input candidate in the considered subset.

This choice was based on the reasonable assumption that if x_t is considered a driver for changes in y, the only other reasonable driver would be x_{t-1}, rather than x_{t-2}.

6. Assuming a relatively fast water exchange between rivers' streambeds and the water table (according to Hatch et al., 2006, river seepage is estimated from daily to monthly), only the last average monthly value of Q is included as an input candidate.

Thus, each well was coupled with the closest discharge station.

7. However, when a streamflow station (or the closest streambed) is located far from the well (or when the exchange of water is too fast to be identified on a monthly scale), Q and GW might be completely uncorrelated. As a consequence, there are two options: whether to include or disregard discharge as an input variable. The latter occurs when the interaction between Q and GW is too fast or when the distance is too long. The former applies when GW and Q are well correlated. Based on the above, if m is the number of candidate subsets resulting from implementing rules 1 to 5, the overall number of candidates to test will be $2m$. Half of them will be the original candidates (i.e., GW and Q are not correlated), and the other half will be the same candidates with the addition of Q_t (i.e. GW and Q are correlated).

After applying the rules above, the total number of input subsets was reduced from 29,791 (if we consider five variables and three lags) to 312. For each combination, the data were divided into training and testing sets; a 10-fold cross-validation was performed on the training set to train an ensemble of 10 multilayer perceptrons (MLPs); and 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.

Also, in order to show the contribution of the exogenous variables on the ANN predicting performances in the different areas of the aquifer, the NSE of the best input combination on the cross-validation set is compared with the NSE obtained by an autoregressive ANN (a neural network forced using only the autoregressive term current-past GW level).

3.3. Modeling Techniques

3.3.1. Artificial Neural Networks

MLP neural networks are a widely used and very well developed technique (Haykin, 2004) in water-related studies (e.g., Elshorbagy et al., 2010b; Abrahart et al., 2012). An MLP is composed of an input layer, a hidden layer, and an output layer. The input layer has as many nodes as the number of inputs, and its nodes just distribute inputs further. The number of nodes in the hidden layer is directly related with the complexity of the problem analyzed and to the number of input neurons. The number of nodes in the output layer is equal to the number of outputs; often, there is only one. The connection between layers occurs through a matrix of weights (w, which have as many rows as the inputs and as many columns as the nodes in the layer), which also expresses the strength of the connection. Nonlinearity is ensured by a sigmoidal transfer function in the nodes of the hidden (and often of the output) layer(s).

When training an MLP, it is important to optimize the number of nodes in the hidden layer. In this study, the number of nodes was selected from a set of values ranging from 5 to 17. The resilient backpropagation algorithm was used to train all neural networks using the R package RSNNS (Bergmeir & Benítez, 2012).

3.4. MultiModel Combination

The ANN models were built for all 300 wells in the aquifer. Predicting performances were evaluated for LTs ranging from 1 to 4 months. For all the prediction horizons, in the case of poorly performing models (NSE < 0.6), the ANN was classified as b. Then, a MuMoC approach using GW predictions from selected neighboring wells was built, with the aim of providing a more accurate forecast. For simplicity and to reduce computational time, we focused on the b wells, and "good" well performances are not further improved.

3.4.1. Selection of Neighboring Wells

Figure 3a illustrates a b well (a black point) characterized as a poorly performing ANN forecast $\widehat{\mathrm{GW}}^{b}_{(t+\tau)}$. The purpose is to develop a model that combines the accurate forecasts $\widehat{\mathrm{GW}}^{ne}_{(t+\tau)}$ from the neighboring ne-wells and ultimately produce an improved forecast. The new forecasted GW level in b will be therefore

$$\widehat{\mathrm{GW}}^{b_{\mathrm{new}}}_{(t+\tau)} = \underset{ne_{\mathrm{s}} \in ne}{f}\left(\widehat{\mathrm{GW}}^{ne_{\mathrm{s}}}_{(t+\tau)}\right) \tag{1}$$

where ne_{s} is the subset of ne, which is used to build the improved model. In fact, among all the possible ne, the selection of neighbors is limited only to those that

1. have accurate forecasts and
2. are well correlated with the observed data in the b-well.

Once the b well is identified, all the ne wells falling within a radius r (optimized by trial and error) are selected (Figure 3b). For each of them, the NSE in the test set NSE(ne) and the cross correlation c(ne,

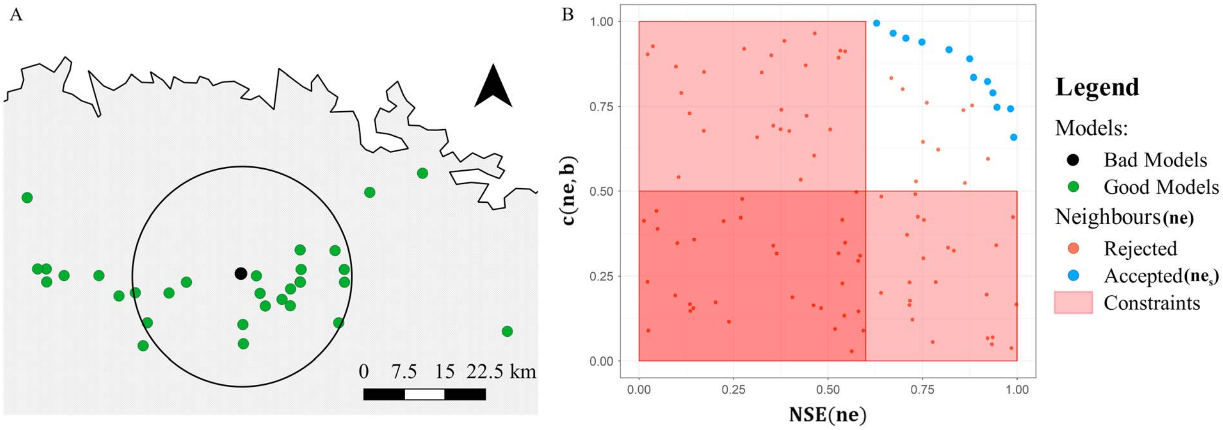

Figure 3. Schematic representation of the process for selecting neighboring wells in MuMoC. (a) Example map representing a *b* well and the *ne* wells in the neighborhood and (b) sketch of a Pareto front in the two-dimensional objective space. MuMoC = MultiModel Combination.

b) with the GW measurements in *b* are computed. Ideally, we would like to select wells having $c(ne, b)$ and $\mathrm{NSE}(ne) \rightarrow 1$. Unfortunately, the two objectives could be conflicting: If $c(ne, b) \rightarrow 1$, then the neighboring model would be similar to *b*, which by definition has low NSE. The best situation one can expect is to have enough neighboring wells that are sufficiently correlated with *b* and that have accurate forecasts at the same time.

Being the two objectives conflicting, ne_s can be found by solving the following optimization problem:

$$ne_s = \underset{ne}{\operatorname{argmax}} |J_1 \;\; J_2| \tag{2a}$$

where

$$J_1 = \mathrm{NSE}(ne) \tag{2b}$$

$$J_2 = c(ne, b) \tag{2c}$$

subject to

$$\mathrm{NSE}(ne) \geq \mathrm{NSE}_{\mathrm{TR}} \tag{2d}$$

$$c(ne, b) \geq c_{\mathrm{TR}} \tag{2e}$$

where J_1 and J_2 are the objective functions to be maximized, $\mathrm{NSE}_{\mathrm{TR}}$ and c_{TR} represent the forecasting accuracy threshold of *ne* and the minimum cross correlation between the *b* and *ne*, for *ne* to be considered a candidate, respectively.

By representing the values assumed by the objectives $\mathrm{NSE}(ne)$ and $c(ne, b)$ in two-dimensional objective function space, like in Figure 3b, it is possible to notice that some of the alternatives might not respect the constraints and are therefore not considered as candidates (orange points in the red rectangles). Among the set of all the *ne* that satisfy the constraints (orange and blue points outside the red rectangles), only those that are not Pareto dominated are selected as input ne_s for MuMoC (blue points). The Pareto front is determined by solving exhaustively the problem defined in equations (2a)–(2e) (the exhaustive search optimization).

3.4.2. Combination of Forecasts at Neighboring Wells

Once the input set for each of the *b* is selected, the ANN forecasts of the ne_s are combined with a *k*-nearest neighbor (*k*-nn) algorithm.

The *k*-nn algorithm belongs to the family of IBL methods. Often, they are also referred to as the "lazy" learning algorithm because, while many machine learning algorithms (such ANN) produce a generalization from the training data as soon as the data have been seen, IBL methods postpone the modeling efforts until a new instance (NI) in the test set becomes available (Witten & Frank, 2005). Once that happens, the NI is compared with existing data using a distance metric (usually the Euclidean distance), and the closest *k* existing

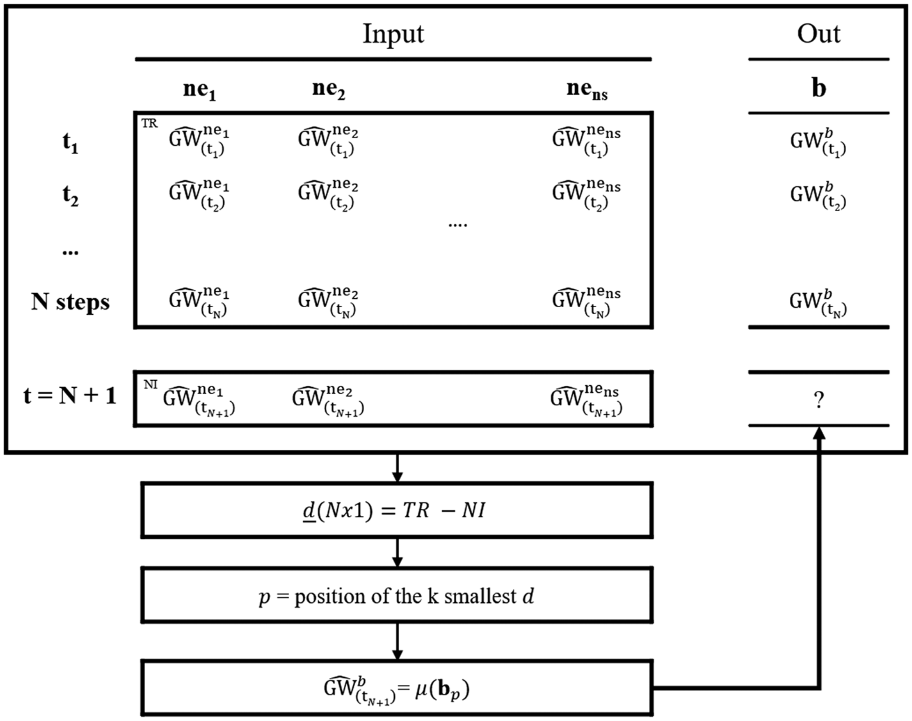
Input
Out
$\mathbf{ne_1}$
$\mathbf{ne_2}$
$\mathbf{ne_{ns}}$
$\mathbf{b}$
TR
$\mathbf{t_1}$
$\mathbf{t_2}$
...
N steps
$\widehat{GW}^{ne_1}_{(t_1)}$ $\widehat{GW}^{ne_2}_{(t_1)}$ $\widehat{GW}^{ne_{ns}}_{(t_1)}$
$\widehat{GW}^{ne_1}_{(t_2)}$ $\widehat{GW}^{ne_2}_{(t_2)}$ $\widehat{GW}^{ne_{ns}}_{(t_2)}$
....
$\widehat{GW}^{ne_1}_{(t_N)}$ $\widehat{GW}^{ne_2}_{(t_N)}$ $\widehat{GW}^{ne_{ns}}_{(t_N)}$
$GW^{b}_{(t_1)}$
$GW^{b}_{(t_2)}$
$GW^{b}_{(t_N)}$
$\mathbf{t = N + 1}$
NI
$\widehat{GW}^{ne_1}_{(t_{N+1})}$ $\widehat{GW}^{ne_2}_{(t_{N+1})}$ $\widehat{GW}^{ne_{ns}}_{(t_{N+1})}$
?
$\underline{d}(Nx1) = TR \; - NI$
p = position of the k smallest d
$\widehat{GW}^{b}_{(t_{N+1})} = \mu(\mathbf{b}_p)$

Figure 4. Schematic representation of forecast combination component in MuMoC: (a) Training matrix and (b) combination algorithm. MuMoC = MultiModel Combination; TR = training matrix; ne_i = ith neighbor; $\left(\widehat{GW}^{ne_i}_{(tj)}\right)$ = forecasted groundwater level in neighbor i at time j; NI = new instance.

instances are used to estimate the output for the new one usually as an average, or the distance-weighted average of the outputs of these closest k instances. This method is also often referred as the k-nn method (Witten & Frank, 2005).

Here the MuMoC method is applied as described in Figure 4 in all the four LTs analyzed in this study: The neighboring wells selection produces an input matrix TR (Figure 4a) with ns columns (as the number of selected neighbors) and N rows (as the number of time steps in the training set). Each instance $(\widehat{\mathrm{GW}}^{i}_{(t_j)})$ in TR represents the ANN GW forecast in the ith neighbor at time step j. When a NI in the test set is available (at time $t = N + j$), the GW level is forecasted with an ANN in each of the neighbor ns (Figure 4B). To estimate $\widehat{\mathrm{GW}}^{b}_{(N+j)}$, the Euclidean distance d between NI and each row of TR is computed. Therefore, d is a row vector with N elements. The k-nn algorithm extracts only the k output instances in TR having the smallest distances and uses them to compute $\widehat{\mathrm{GW}}^{b}_{(N+j)}$ as the average of their corresponding outputs.

To implement the model combination, we chose four parameters: $\mathrm{NSE_{TR}}$ (NSE threshold), c_{TR} (cross-correlation threshold), r (radius of influence), and k (number of neighbors). The value of $\mathrm{NSE_{TR}}$ and c_{TR} are set to 0.6 and 0.5, respectively; r was selected in each well by trial and error: The radius formed between a b well and a neighboring was progressively increased by 10 km until no improvements in the Pareto set were found for three consecutive progressions. The number of neighbors k was selected from a set of values ranging from 2 to 11 by minimizing the error on the cross-validation set. The k-nn algorithm was implemented using the R package *lazy* (Birattari & Bontempi, 2003).

3.5. Metrics of Performance

Metrics of performance are used to express the skill of the forecast by aggregating model residual in time (Fenicia et al., 2007). Since different metrics may enhance different aspects of the error while neglecting others, the use of multiple error statistics is recommended (Gupta et al., 1998). To this aim, we use three metrics of performance: the NSE, the coefficient of determination (R^2), and the normalized RMSE (RMSE_p). All estimated for the 300 observation wells. The main reason behind the choice of the three error statistics is to identify: (1) the value of the error with respect to the variance (NSE), (2) the correlation between the observed and predicted value (R^2); and (3) the value of the squared residuals with respect to the average (RMSE_p). The NSE (equation (3a)) provides a score in the interval $(-\infty; 1]$ for the error variance. An NSE value equal to 1 represents a perfect predictor, while an NSE value equal to 0 represents the predicting capability of the average of the population. R^2 (equation (3b)) indicates the strength of the correlation between observed and predicted values. It can vary between 0 and 1, with values close to the unity, indicating strong correlation. The RMSE provides insights into the square difference between the observed observations and simulation. Here, to facilitate the comparison between different scales, the RMSE is normalized by the average (equation (3c)).

$$\mathrm{NSE} = 1 - \frac{\sum_{i=1}^{N}(O_i - P_i)^2}{\sum_{i=1}^{N}(\overline{O} - O_i)^2} \tag{3a}$$

$$R^2 = \left(\frac{\sum_{i=1}^{N}(O - \overline{O})(P - \overline{P})}{\sqrt{\sum_{i=1}^{N}(\overline{O} - O_i)^2}\sqrt{\sum_{i=1}^{N}(\overline{P} - P_i)^2}}\right)^2 \tag{3b}$$

$$\mathrm{RMSE}_p = \frac{\sqrt{\frac{\sum_{i=1}^{N}(O_i - P_i)^2}{N}}}{\overline{O}} \tag{3c}$$

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

To assess the predictive capability of the ANN at different locations and at different LTs, we study the variation of the three indexes both spatially (different wells) and temporally (for an increase in LTs).

To quantify the improvement in predicting capabilities brought by implementing MuMoC in b, we compare (in all the b) the NSE obtained by MuMoC in the test set with the one obtained by the single ANN. Mathematically, this can be expressed as

$$\Delta\mathrm{NSE} = \mathrm{NSE}_{\mathrm{MuMoC}} - \mathrm{NSE}_{\mathrm{ANN}} \tag{3d}$$

4. Results and Discussion

4.1. Evaluation of ANN Forecasting Capability Across the HP Aquifer

The modeling experiments produced the results presented in Figure 5, which represents the NSE in the testing set for the four LTs analyzed in this study. Looking at this figure and Table 3, one can observe overall good modeling performances across the aquifer. This is particularly true for an LT of 1 month, in which situation only Kansas has an NSE lower than 0.9.

In Figure 5, one can notice an increase in the error (decrease in NSE) when the LT increases. However, this did not happen uniformly throughout the aquifer: The decrease in NSE between 1 and 4-month LTs was marginal in the southern states of Texas (0.03), New Mexico (0.002), and Oklahoma (0.05) and in the northern state of South Dakota (0.07); on the other hand, in Nebraska (the eastern part in particular) and Kansas, the decrease in model performance was more evident as the LT increased. By looking at Table 3, one can see that the NSE value decreased from 0.85 (LT = 1 month) to 0.72 (LT = 4 months) in Kansas and from 0.93 (LT = 1 month) to 0.81 (LT = 4 months) in Nebraska.

The explanation for this might lie in the different hydrogeological and land use conditions that occur through the aquifer. In the southern area (from Texas up to the Kansas-Oklahoma border), the aquifer is usually much deeper (with depth peaks of more than 100 m below land surface) than in the north. Since the hydraulic conductivity through the aquifer is approximately in the same order of magnitude (10^{-3}–10^{-4} m/s in more than 95% of the HP; Gutentag et al., 1984), high depth causes here a delay in the contribution of meteorological variables, precipitation in particular (which is already less than half that in eastern Nebraska and eastern Kansas). In addition, due to the high difference in elevation between the riverbed and the water table level (and the absence of major streams in the area), the interaction between surface water and GW in this area can also be considered null.

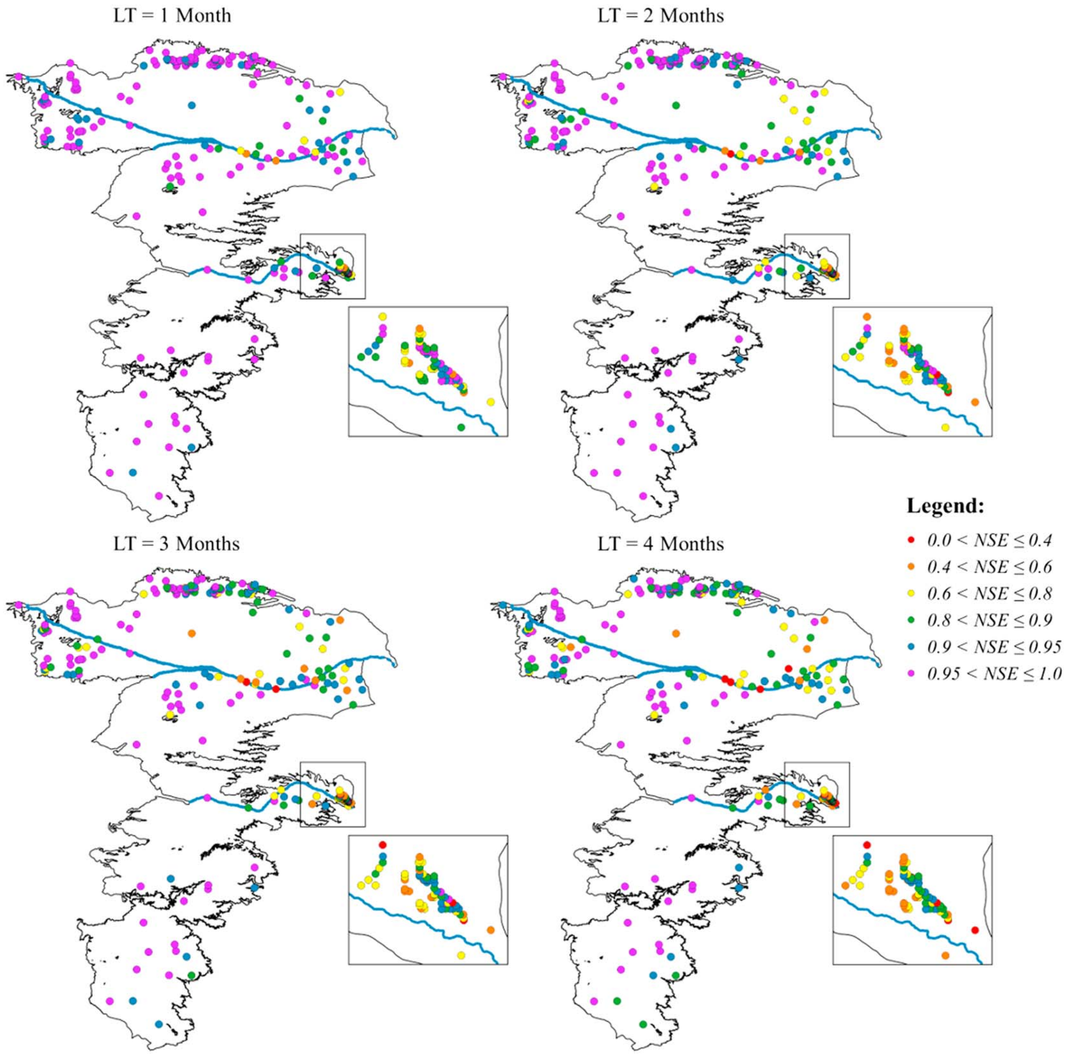

Figure 5. Nash-Sutcliffe efficiency (NSE) in the testing set in the four lead times (LTs) analyzed.

Figure 6 shows the R^2 and the RMSE$_p$ values (at 4 months' LT) with respect to the average GW table depth. Here a high Pearson correlation coefficient (0.96) and low error values (RMSE$_p$ in the order of 10^{-3}, corresponding to a RMSE of about 10 cm) in areas with high GW depth seem to indicate that the southern portion of the aquifer is nonsensitive to hydroclimatic forcings. The main drivers of water table changes are GW withdrawals for irrigation. Consequently, GW levels in this area experienced a slow depletion trend in the past 60 years (as confirmed by McGuire, 2017), and its dynamics are dominated by the autoregressive component. This is also supported by the high autocorrelation of the GW level time series (usually higher than 0.85 after 4 months) and by the CIVS results. In this regard, Figure 7 shows the improvement in performances (in terms of NSE) when exogenous variables are included in the input set. As can be seen in Figure 7, in the southern part of the aquifer the improvement is marginal (often below 0.05 in NSE). The presence of the dominant autoregressive component creates the perfect condition for obtaining a very high forecast performance even with high LTs. A good modeling performance in deep and relatively isolated aquifers (with respect to shallow ones) was also found by Rakhshandehroo et al. (2018).

Table 3
Average Nash-Sutcliffe Efficiency (NSE) Values in the Testing Set

State	*N* wells	μNSE			
		1 month	2 months	3 months	4 months
South Dakota	70	0.97	0.95	0.93	0.90
Nebraska	68	0.93	0.88	0.84	0.81
Colorado	1	0.99	0.99	0.99	0.99
Kansas	110	0.85	0.79	0.75	0.72
Wyoming	32	0.97	0.94	0.93	0.91
New Mexico	1	0.99	0.99	0.99	0.99
Oklahoma	2	0.98	0.95	0.95	0.93
Texas	15	0.98	0.98	0.97	0.95

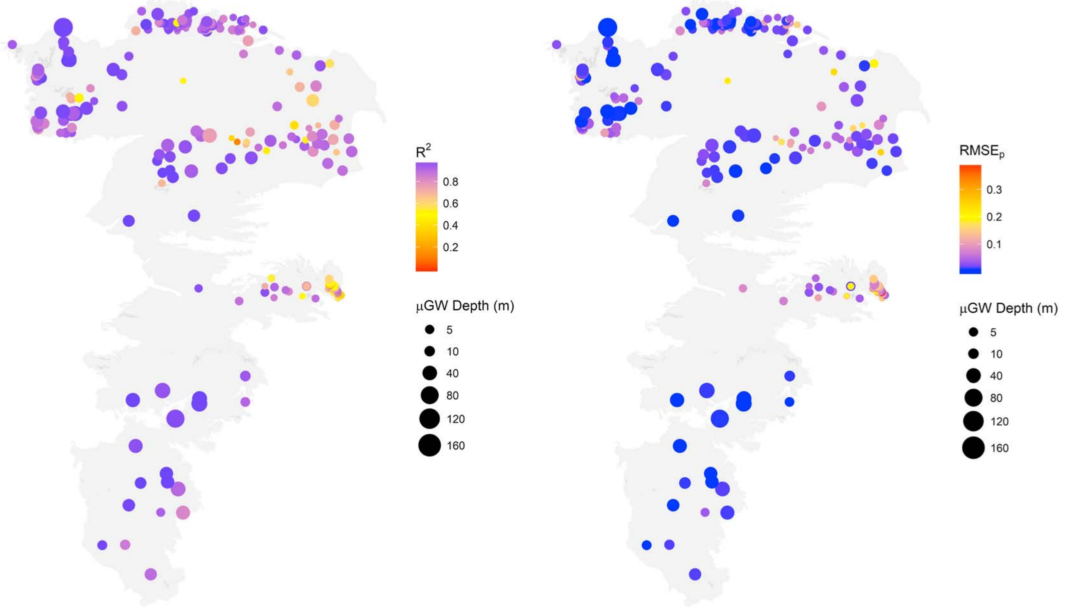

Figure 6. Graphical representation of (left) R^2 and (right) RMSE$_p$ values versus average groundwater (GW) level in the testing set for 4 months' LT.

In the northern part of the aquifer (South Dakota), the water level is shallow (usually between 5 and 20 m below the land surface, as can be seen from Figure 6). However, the recharge is small (from less than half to less than one tenth of Nebraska's, for example), the hydraulic conductivity is the lowest in the aquifer (10^{-4} according to Gutentag et al., 1984; 10^{-6} according to Luo & Pederson, 2012), and the irrigation intensity in that area is also low (less than 20%; Figure 1d) because most of the cropping lands are rain fed. This may lead to a delay in the aquifer's recharge from precipitation, since a good fraction of precipitation might be lost to leaf interception and another fraction to fulfilling the crop WD. This is also observed in negligible interseasonal variability in GW table changes.

In contrast, a significant decrease in model performances is observed in eastern Nebraska and Kansas, especially along the Platte and Arkansas Rivers, where the NSE and R^2 values are sometimes lower than 0.6 and the RMSE$_p$ reaches its peak (0.36).

Eastern Nebraska is the most intensively cultivated area in the aquifer (as indicated by the predominant red color in this particular area of Figure 1d), with GW-based irrigation intensity often higher than 90%. In addition, here the estimated net recharge to the aquifer is around 150 mm/year (Scanlon et al., 2012), about 100 times the estimated recharge value for Texas. A shallow water table (most of the time lower than 10 m below surface) enables a strong and fast interaction between surface water and GW. Thus, high water consumption from irrigation and high ET rates cause major and fast water depletions during the growing season. However, high rainfall (as can be seen from Figure 1c, eastern Nebraska has, together with eastern Kansas, the highest rainfall in the aquifer) and a strong interaction with streamflow bring the water table levels back to long-term stationary conditions as soon as the growing season ends. Analysis of the CIVS results in this area revealed, as shown in Figure 7, an improvement in the model performance based on increases in NSE of about 0.23 when exogenous variables were included in the input set with respect to an autoregressive neural network. Therefore, the combination of those land use, climatic, and hydrogeological conditions contribute to the nonlinearity of the system and, consequently, affect the performance of the model. Similar results were also found by Amaranto et al. (2018), who illustrate how the fast late-spring-to-summer withdrawal and autumn-to-early-spring recharge cause fast and difficult-to-predict water level changes. The same occurs in eastern Kansas (along the Arkansas River). Here despite the influence of irrigation intensity and precipitation-driven recharge being lower than those in western Nebraska, they are much higher than those wells in Texas and South Dakota (as can be seen in Figures 1c and 1d). The

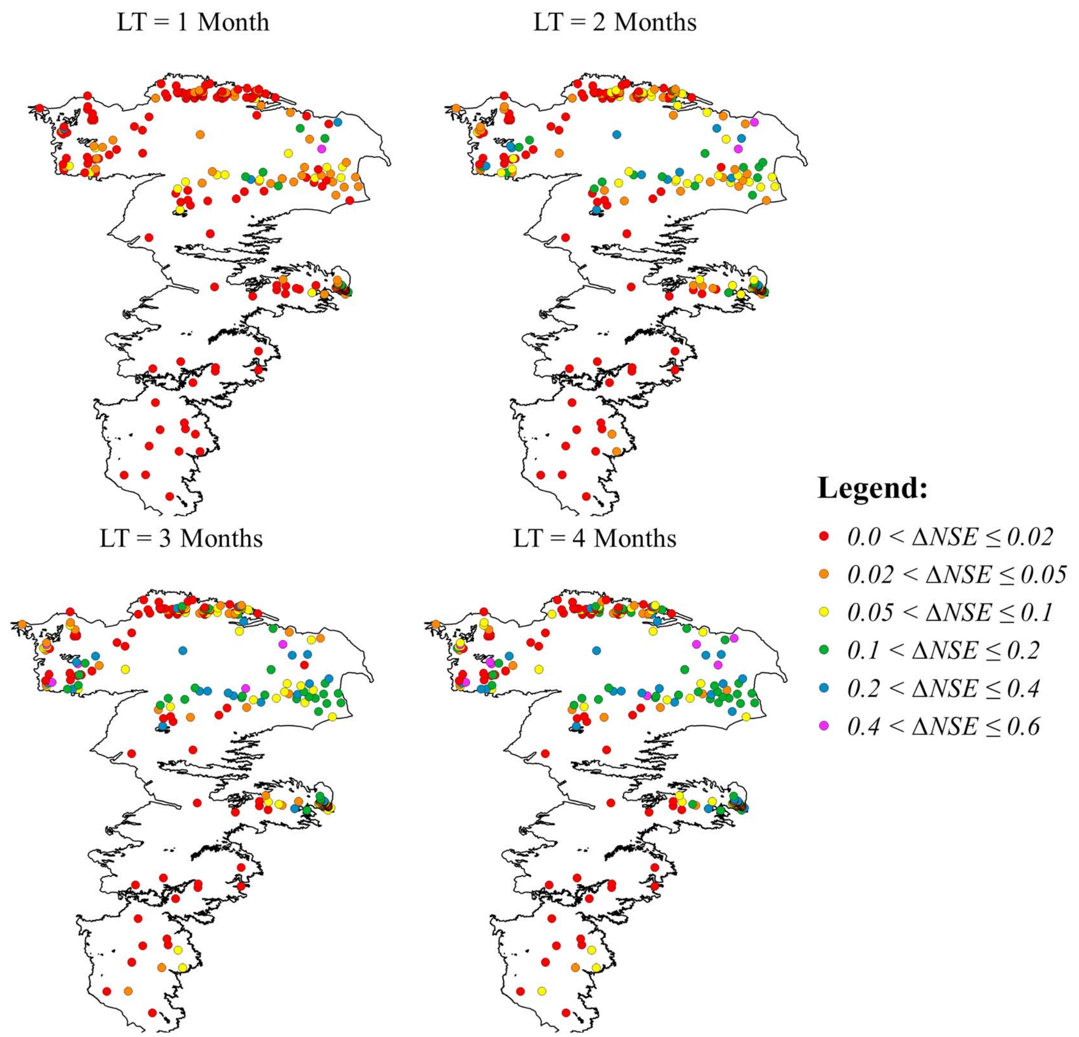

Figure 7. ΔNSE values obtained in the four lead times (LTs) under investigation. ΔNSE is defined as the difference in Nash-Sutcliffe Efficiency (NSE) obtained in the cross-validation set between the artificial neural network generated by the CIVS and an autoregressive artificial neural network with no exogenous input.

interdependence between precipitation and irrigation with the GW level changes is also evident in the Arkansas River basin, where the GW-Arkansas River interactions confirmed what Scanlon et al. (2012) reported.

In summary, areas of Nebraska and eastern Kansas are influenced by climate and management and regulated by surface water-GW interactions. Those interdependencies bring high nonlinearities into the system, decreasing the GW autocorrelation (very often below 0.4 after 2 months) and, combined with the unavailability of observed pumping data, causing a loss in forecasting accuracy.

4.2. Evaluation of MuMoC Performances

Figure 8 shows the locations of the wells for which predicting performances did not satisfy the minimum NSE requirement (NSE > 0.6). Figure 8 also evidences that all the wells are located either in Kansas or in Nebraska. For this reason, further analysis will be devoted to those two states.

As can be seen from Table 4, most of the *b* wells are in Kansas. Furthermore, more sparsely distributed observations in Nebraska led to an average lower number of neighboring wells selected (ne_S). In two locations at LTs 3 and 4 months it was practically impossible to find any neighboring wells that addressed all the constraints presented in equations (2b) and (2c) (and they were therefore excluded from further analysis).

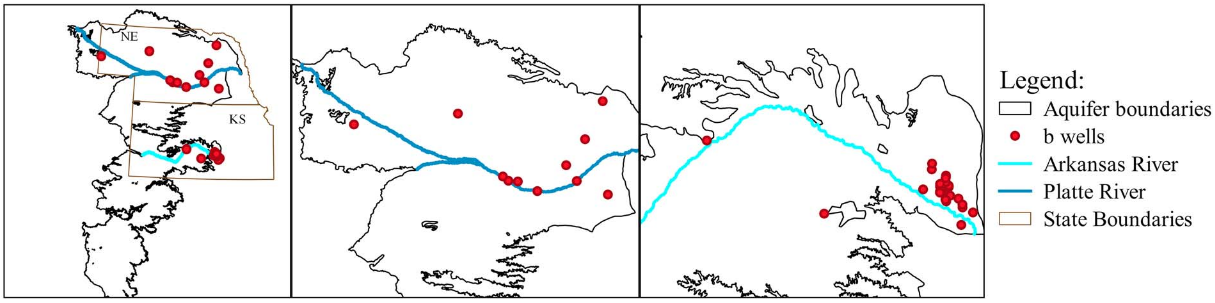

Figure 8. Locations of *b* wells.

The improvement of the model's performance is evident by comparing Figure 9a with 9b. Implementing the MuMoC brought the average NSE in Kansas very close to the 0.6 threshold for the four LTs analyzed and brought an overall average improvement in the performance of about 25%. As a result, about 50% of the wells in Kansas shifted from *b* to the good status. Results for Nebraska's *b* wells were less encouraging, since the values of the NSE for MuMoC were lower and the improvement was negligible (ΔNSE = 0.02). One explanation for this might be the lower number of neighboring wells available. Also, the variation of the GW level may change dramatically even in a small region due to the heterogeneity of soil and rock medium. It is therefore noteworthy to mention that the dense monitoring well network in Kansas allowed for the identification of neighbors within 3 km from *b*. The cross-correlation values (often higher than 0.75 in this particular area) led the authors to believe that the conditions inherent to the aquifer system for *b* and ne_s might be similar. This is also supported by the fact that the area in Kansas under investigation (unlike the remaining 75% of the aquifer) belongs to the Quaternary undivided formation (Gutentag et al., 1984). It is therefore possible to observe here high intraneighbor similarity. In addition, the values shown in Figure 9 are an aggregated statistic that summarizes results by state. It is noteworthy that improvement in model performances can vary from well to well and from region to region. Such variation on a smaller scale (evident in the differences across states) might be explained by the quality of the selected neighbors in Nebraska, where ne_s wells were almost always less correlated with *b* and had lower predicting accuracy when compared with Kansas.

Based on the above, Figure 10 shows the improvement in forecasting accuracy derived from the implementation of MuMoC with respect to ANN (ΔNSE, color of the circles) and the forecasting accuracy of MuMoC ($\text{NSE}_{\text{MuMoC}}$, size of the circles). The *x*-axis represents the forecasting accuracy in the optimal neighbor ne_s, while the *y*-axis represents the cross correlation between the optimal neighbor and the *b* well. The optimal ne_s is the neighbor that minimizes the Euclidean distance from the ideal point $(\max[\text{NSE}(ne_s), c(ne_s, b)])$. Strong improvement combined with the good modeling performance is indicated by a big red circle in the two-dimensional objective space.

Figure 10 illustrates how the MuMoC performance and performance improvement are strongly related with the overall neighbors' data quality. In fact, when $\text{NSE}(ne_s)$ and $c(ne_s, b)$ are lower than 0.75 and 0.7, respectively, MuMoC performances are always lower than those based just on ANNs (with a minimum ΔNSE of −0.2 and average NSE of 0.4). This case is relevant to Nebraska's *b* wells (nine out of 68 cases). On the other hand, when $\text{NSE}(ne_s)$ and $c(ne_s, b)$ are higher than 0.9 and 0.75, respectively, there is an average ΔNSE of about 0.18, with improvements of maximum performance (ΔNSE) of about 0.30. In addition, all the wells that belong to this category shifted from the *b* condition to the good condition, with an average NSE value of 0.67 and maximum of 0.77. The main reason for such a strong improvement might be that the spatial correlation of the GW variations in the neighborhood of *b* dominated the temporal autocorrelation component, which in the case of the *b* wells was particularly low (often lower than 0.4 after 3 months). Therefore, when the forecasts in ne_s are also accurate enough, MuMoC represents a more robust approach leading to an increase in the semiseasonal forecast skill with respect to ANNs.

Table 4
Number of b Wells (Nb) and Average Number of Selected Neighbors $\overline{N}ne_s$ per b in Kansas and Nebraska From 1 to 4 Months' Lead Time (LT)

	LT							
	1 month		2 months		3 months		4 months	
State	*Nb*	$\overline{N}ne_s$	*Nb*	$\overline{N}ne_s$	*Nb*	$\overline{N}ne_s$	*Nb*	$\overline{N}ne_s$
Nebraska	2	3.5	4	2	9	1.3	9	2
Kansas	8	5	17	5.8	19	5.3	29	5

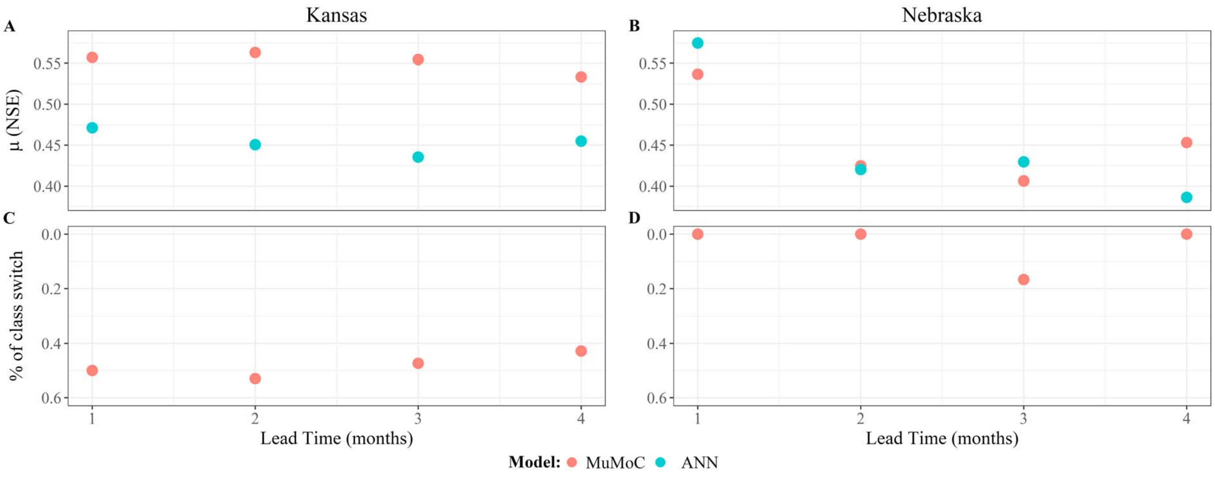

Figure 9. (top row) Comparison between the NSE for the ANN and MuMoC approach by state: (a) Kansas and (b) Nebraska. (bottom row) Fraction of wells that shifted from the *b* to good class following the implementation of MuMoC: (c) Kansas and (d) Nebraska. NSE = Nash-Sutcliffe efficiency; ANN = artificial neural network; MuMoC = MultiModel Combination.

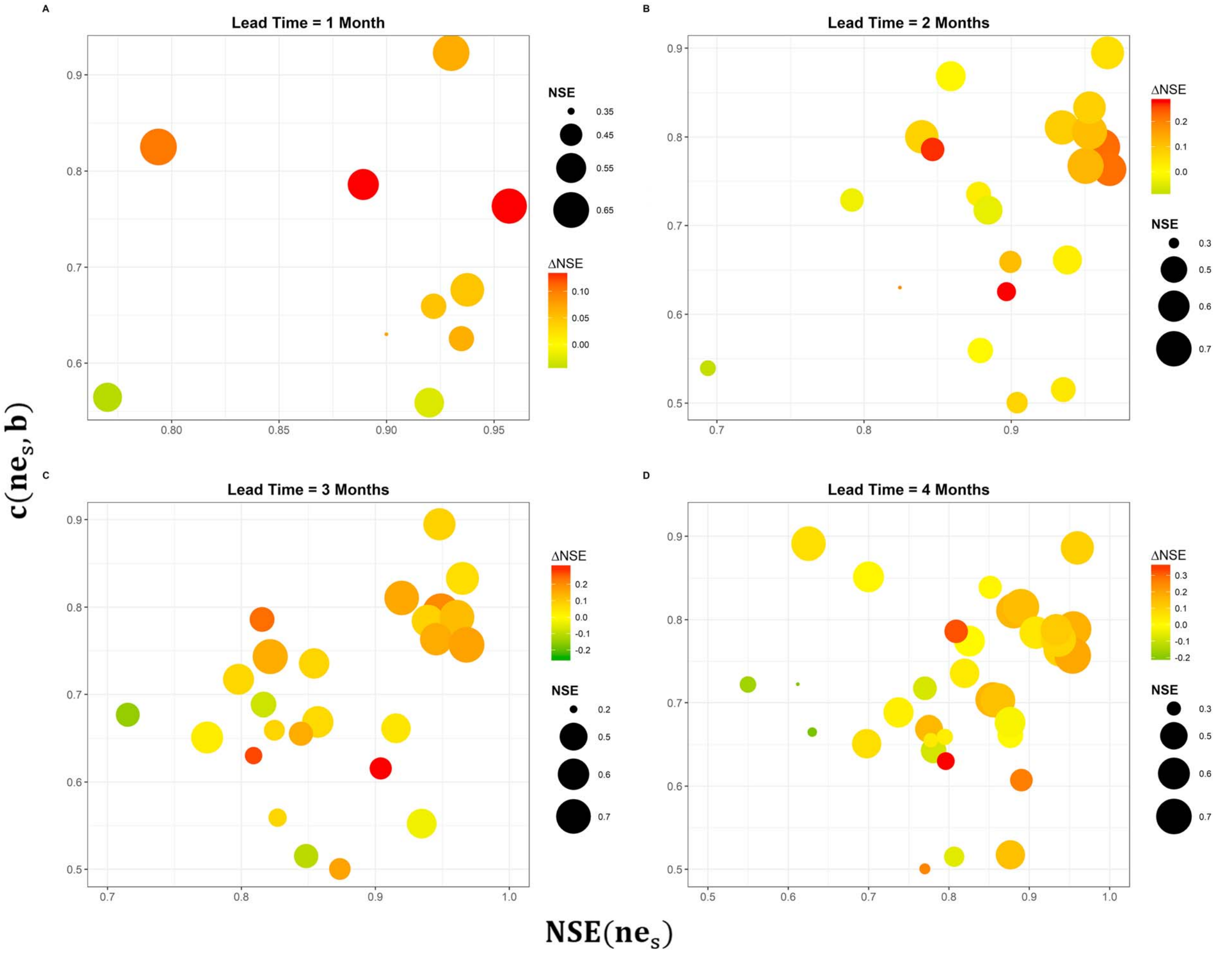

Figure 10. MuMoC NSE (size of the circle) and ΔNSE ($NSE_{MuMoc} - NSE_{ANN}$, color of the circle) in the *b* wells for LTs of (a) 1 month, (b) 2 months, (c) 3 months, and (d) 4 months. *c* (ne_s, *b*) is the cross correlation between *b* and the optimal neighbor. NSE (ne_s) is the NSE in the neighbor. MuMoC = MultiModel Combination; NSE = Nash-Sutcliffe efficiency; LT = lead time.

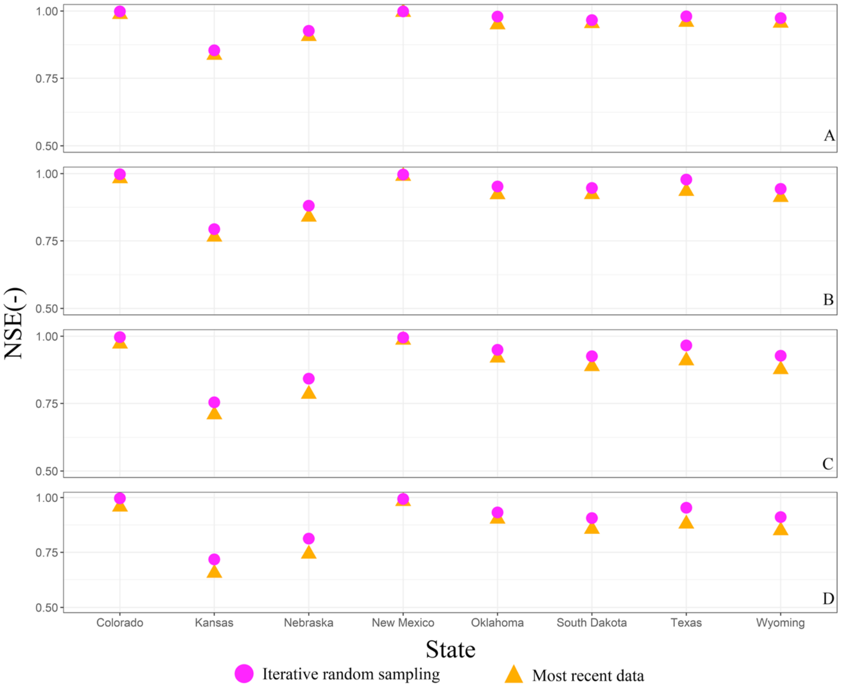

Figure 11. Statewise comparison of ANN performances when iterative random sampling or most recent data are used as test set: (a) 1-month lead time; (b) 2-month lead time; (c) 3-month lead time; and (d) 4-month lead time. ANN = artificial neural network; NSE = Nash-Sutcliffe efficiency.

4.3. Performances Evaluation on the Most Recent Data

Figure 11 represents the results obtained by testing the ANN on the most recent data (final 30% of the time series), in comparison with those achieved by iteratively splitting the training and test set (to maximize the statistical similarity between them). Also, Figure 11 evidences that changing the splitting procedure leads to imperceptible variations in forecasting performance in all the eight states and all the four LTs under analysis. The error is marginally lower when the statistical similarity between the training and the test set is maximized. This result was expected, since the criteria for the training set selection ensured that the models were calibrated on a range of values very similar to those used in testing. However, the maximum difference in terms of forecasting accuracy (occurred in Nebraska for a LT of 4 months) does not go beyond −0.06 in NSE, with an average decrease in NSE of about 4% when the model is tested on the most recent data.

A similar pattern can be observed when analyzing MuMoC performances (see Table 5).

As occurred in the previous analysis, all the wells that did not satisfy the constraints were located in either Kansas or Nebraska, but the marginal decrease in ANN performance increased the overall number of *b* wells (from 10 to 12 when LT = 1 month; from 38 to 43 when LT = 4 months).

Table 5
MuMoC Results When Tested on the Most Recent Data

Lead time	Nb	$\mu\text{NSE}_{\text{MuMoC}}$	$\mu\Delta\text{NSE}$	max(ΔNSE)	% class shift
1 month	12	0.48	0.11	0.35	0.36
2 months	24	0.45	0.14	0.41	0.35
3 months	32	0.44	0.13	0.33	0.40
4 months	43	0.48	0.12	0.38	0.40

Note. Nb = number of *b* wells; $\mu\text{NSE}_{\text{MuMoC}}$ = average NSE obtained with MuMoC; $\mu\Delta\text{NSE}$ = average performance improvement with respect to ANN; max(ΔNSE) = maximum improvement; ANN = artificial neural network; NSE = Nash-Sutcliffe efficiency; MuMoC = MultiModel Combination.

When tested on the most recent data, MuMoC again outperformed ANN in all the LTs under investigation, with an average performance improvement of about 0.12 in NSE (23% improvement), a maximum improvement of 0.41 in NSE, and an average of 38% of the wells shifting from the bad to good class.

Furthermore, by comparing the numerical values in Table 5 with those in Table 4 and Figures 9 and 10, an overall similitude is noticed between modeling performances with those obtained with the iterative random sampling for the test set selection. The explanation for such a small variation when changing the test set (for both ANN and MuMoC) probably lies in the fact that the statistical properties (average and standard deviation) of the GW time series does not change dramatically through time. This allows the training and the test set to fit approximately the same statistical distribution (even if not to the similarity extent obtained with the iterative random sampling), and the ANN to be able to extrapolate the NIs with only a marginal loss of accuracy with respect to those when the split is optimized.

5. Conclusions

The goal of this paper was to develop a GW forecasting framework to improve the semiseasonal (1- to 4-month) predictability of water level changes, at the aquifer scale in the HP (USA). A data-driven modeling approach based on ANNs was used to forecast semiseasonal GW changes in 300 wells across the HP aquifer. ANN forecasting abilities were evaluated using NSE, R^2, and RMSE$_p$. Furthermore, the values of the error statistics are contextualized with different hydrogeological, land use, and meteorological conditions across the aquifer (objective 1). Then, when ANN performance did not satisfy the minimum NSE requirement (0.6) for a certain well, we proposed an alternative modeling framework named MuMoC. MuMoC includes an optimization algorithm to select wells in the geographical neighborhood of a bad (*b*) performing model (using cross correlation and NSE in the neighbors as optimization criteria to maximize). Then, by combining the ANN-forecasted GW level in the neighborhood, MuMoC provided an updated (and an expected improved forecast) value in *b* (objective 2). The implementation of the proposed framework evidenced the following:

1. Overall, the single-model (ANN) approach exhibited high forecasting accuracy through all the aquifer. The average NSE was higher than 0.8 even when the LT was increased to 4 months in all the states.
2. The expected decrease in predictability of GW well levels using ANN with respect to LT was more conspicuous in eastern Nebraska (−0.12 in NSE from LT = 1 to LT = 4) and Kansas (−0.13 in NSE from LT = 1 to LT = 4), probably due to the strong effect that the integrated hydrometeorological and management components have on the GW systems in those areas. Here precipitation had a strong influence, combined with high surface-GW interaction and high irrigation intensity. This finding was also supported by the IVS results, which showed an average 0.2 decrease in NSE when exogenous variables were excluded from the input set.
3. Decreases in performance with increasing LT were negligible in the southern part of the aquifer (~0.02) where the GW system was strongly autocorrelated and the influence of exogenous variables was negligible. The constrained contribution of irrigation, recharge rates, deep aquifer water tables, and negligible surface water-GW interaction requires further investigation to determine the causality of such good forecasting predictability.
4. Overall, MuMoC improved semiseasonal forecasts of changes in GW well levels by about 25%, based on the NSE values of bad wells mainly in Kansas. However, the lack of good wells geolocated near the bad wells in Nebraska led to poor improvements on the 1- to 4-month LTs. Conclusion number 2 may help illustrate the complexity and causality of changes in GW well levels and also points up the need for additional data (a higher sampling frequency) of wells with shallow water tables and strong surface-GW interactions.
5. The improvement in performance brought by implementing MuMoC proved particularly sensitive to the quality of neighboring wells. When neighboring wells showed a high correlation (>0.75) with *b* and good forecasting capability (>0.9), all the wells shifted to the good class (NSE from <0.6 to >0.6), with an average improvement (ΔNSE) of 0.18 (with peaks of ΔNSE = 0.3) and an average NSE of 0.67 (with peaks of NSE = 0.77).

This study had three limitations. First, the feasibility of implementing MuMoC depends on the presence of neighboring wells and, therefore, on data availability. Consequently, as is true with many models that include a geospatial component, the method cannot be applied in very sparsely gauged aquifers where only two or three stations are available. In addition, it is noteworthy to mention that the performances of MuMoC were strongly dependent not just on the neighbor's availability but also on their qualities: forecasting accuracy and spatial correlation. This implies that an application of MuMoC is recommended in densely gauged areas, in such a way that neighbors having similar physical trajectories (i.e., high correlations with the *b* wells) of GW level are more likely to be found. We also encourage as a future research direction the testing of MuMoC on different aquifers with different observation density across the globe, in order to further understand its range of applicability. Second, the absence of observed pumping data limits this study. Even though crop WD and ET are used as proxies, the ANN accuracy is lower when irrigation is higher. We recommend directing future research toward identifying better proxies to represent unknown pumping patterns, perhaps by using remote sensing estimated ET trends and studying the trade-off between less data availability (Moderate Resolution Imaging Spectroradiometer is available only for the past 18 years) and improved pumping proxy. Third, the influence of the exogenous variables on the forecasting skills of ANN is assessed here without distinguishing the specific contribution of each input. Consequently, we suggest for a future research direction to implement a global sensitivity analysis of ANN models to observational endogenous and exogenous inputs uncertainties. Sensitivity analysis would contribute toward understanding the spatial distribution of the dominant principles of the hydrogeological processes across the aquifer and guide the modelers toward a more robust decision of the variables to be selected. This could be done by employing state-of-the-art techniques of sensitivity analysis (see, e.g., Pianosi et al., 2016).

In general, the results of this study were encouraging, indicating a forecasting framework that, when implemented in operational practice, practitioners may use to improve GW management in the HP.

Acknowledgments

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 1009760 Grant 17-77-30006 of the Russian Science Foundation, and the project Irrigation Sustainability funded by the University of Nebraska-Lincoln's Agriculture Research Division. Post-processed input data and experiments results can be founded in https://github.com/alessandroamaranto/MuMoC. Finally, we appreciate the valuable comments and suggestions made by the reviewers and Maurizio Mazzoleni.

References

Abrahart, R. J., Anctil, F., Coulibaly, P., Dawson, C. W., Mount, N. J., See, L. M., et al. (2012). Two decades of anarchy? Emerging themes and outstanding challenges for neural network river forecasting. *Progress in Physical Geography: Earth and Environment*, *36*(4), 480–513. https://doi.org/10.1177/0309133312444943

Allen, R. G., Pereira, L. S., Raes, D., & Smith, M. (1998). Crop evapotranspiration—Guidelines for computing crop water requirements (FAO Irrigation and Drainage Paper 56). FAO, Rome, 300(9), D05109.

Amaranto, A., Munoz-Arriola, F., Corzo, G., Solomatine, D. P., & Meyer, G. (2018). Semi-seasonal groundwater forecast using multiple data-driven models in an irrigated cropland. *Journal of Hydroinformatics*, jh2018002.

Bartolino, J., & Cunningham, W. (2003). Ground-water depletion across the nation. *Drinking Water and Backflow Prevention*, *29*(1), 26–29.

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.

Bhattacharya, B., & Solomatine, D. P. (2006). Machine learning in sedimentation modelling. *Neural Networks*, *19*(2), 208–214. https://doi.org/10.1016/j.neunet.2006.01.007

Birattari, M., & Bontempi, G. (2003). lazy: Lazy learning for local regression, 2003. R package version, 1.2-14.

Blaney, H. F., & Criddle, W. D. (1962). *Determining Consumptive Use and Irrigation Water Requirements*, *Technical Bulletin*, (Vol. 1275). Washington, DC: U.S. Department of Agriculture.

Bowden, G. J., Maier, H. R., & Dandy, G. C. (2005). Input determination for neural network models in water resources applications. Part 2. Case study: Forecasting salinity in a river. *Journal of Hydrology*, *301*(1-4), 93–107. https://doi.org/10.1016/j.jhydrol.2004.06.020

Butler, J. J. Jr., Whittemore, D. O., Wilson, B. B., & Bohling, G. C. (2018). Sustainability of aquifers supporting irrigated agriculture: A case study of the High Plains aquifer in Kansas. *Water International*, *43*(6), 815–828. https://doi.org/10.1080/02508060.2018.1515566

Christiansen, D. E. (2012). *Simulation of Daily Streamflows at Gaged and Ungaged Locations Within the Cedar River Basin, Iowa, Using a Precipitation-Runoff Modeling System Model*, *Scientific Investigations Report 2012-5213*. Reston, VA: U.S. Geological Survey.

Coppola, E. Jr., Szidarovszky, F., Poulton, M., & Charles, E. (2003). 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. https://doi.org/10.1061/(ASCE)1084-0699(2003)8:6(348)

Coppola, E. A. Jr., Rana, A. J., Poulton, M. M., Szidarovszky, F., & Uhl, V. W. (2005). A neural network model for predicting aquifer water level elevations. *Groundwater*, *43*(2), 231–241. https://doi.org/10.1111/j.1745-6584.2005.0003.x

Dalin, C., Wada, Y., Kastner, T., & Puma, M. J. (2018). Corrigendum: Groundwater depletion embedded in international food trade. *Nature*, *553*(7688), 366. https://doi.org/10.1038/nature24664

di Baldassarre, G., Viglione, A., Carr, G., Kuil, L., Salinas, J., & Blöschl, G. (2013). Socio-hydrology: https://doi.org/g human-flood interactions. *Hydrology and Earth System Sciences*, *17*(8), 3295–3303. https://doi.org/10.5194/hess-17-3295-2013

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. https://doi.org/10.5194/hess-14-1931-2010

Elshorbagy, A., Corzo, G., Srinivasulu, S., & Solomatine, D. P. (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. https://doi.org/10.5194/hess-14-1943-2010

Fenicia, F., Savenije, H. H., Matgen, P., & Pfister, L. (2007). A comparison ofalternative multiobjective calibration strategies for hydrological modeling. *Water Resources Research*, *43*, W03434. https://doi.org/10.1029/2006WR005098

Galelli, S., & Castelletti, A. (2013). Tree-based iterative input variable selection for hydrological modeling. *Water Resources Research*, *49*, 4295–4310. https://doi.org/10.1002/wrcr.20339

Galelli, S., Gandolfi, C., Soncini-Sessa, R., & Agostani, D. (2010). Building a metamodel of an irrigation district distributed-parameter model. *Agricultural Water Management*, *97*(2), 187–200. https://doi.org/10.1016/j.agwat.2009.09.007

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. https://doi.org/10.1016/j.envsoft.2014.08.015

Giuliani, M., Castelletti, A., Pianosi, F., Mason, E., & Reed, P. M. (2015). Curses, tradeoffs, and scalable management: Advancing evolutionary multiobjective direct policy search to improve water reservoir operations. *Journal of Water Resources Planning and Management*, *142*(2), 04015050. https://doi.org/10.1061/(ASCE)WR.1943-5452.0000570

Giuliani, M., Mason, E., Castelletti, A., Pianosi, F., & Soncini-Sessa, R. (2014). Universal approximators for direct policy search in multipurpose water reservoir management: A comparative analysis. *IFAC Proceedings Volumes*, *47*(3), 6234–6239. https://doi.org/10.3182/20140824-6-ZA-1003.01962

Gupta, H. V., Sorooshian, S., & Yapo, P. O. (1998). Toward improved calibration of hydrologic models: Multiple and noncommesurable measures of information. *Water Resources Research*, *34*(4), 751–763. https://doi.org/10.1029/97WR03495

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*, *USGS Professional Paper 1400-B*. Washington, DC: U.S. Government Printing Office.

Guzman, S. M., Paz, J. O., & Tagert, M. L. M. (2017). The use of NARX neural networks to forecast daily groundwater levels. *Water Resources Management*, *31*(5), 1591–1603. https://doi.org/10.1007/s11269-017-1598-5

Hatch, C. E., Fisher, A. T., Revenaugh, J. S., Constantz, J., & Ruehl, C. (2006). Quantifying surface water–groundwater interactions using time series analysis of streambed thermal records: Method development. *Water Resources Research*, *42*, W10410. https://doi.org/10.1029/2005WR004787

Haykin, S. (2004). *Neural Networks: A Comprehensive Foundation*. Prentice Hall, Upper Saddle River: N.J. Macmillan College Publishing.

Houston, N. A., Gonzales-Bradford, S. L., Flynn, A. T., Qi, S. L., Peterson, S. M., Stanton, J. S., et al. (2013). *Geodatabase Compilation of Hydrogeologic, Remote Sensing, and Water-Budget-Component Data for the High Plains Aquifer, 2011*, Data Series, (Vol. *777*). Reston, VA: U.S. Geological Survey.

Iglesias, A., & Garrote, L. (2015). Adaptation strategies for agricultural water management under climate change in Europe. *Agricultural Water Management*, *155*, 113–124. https://doi.org/10.1016/j.agwat.2015.03.014

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. https://doi.org/10.1016/j.proeng.2016.07.528

Luo, W., & Pederson, D. T. (2012). Hydraulic conductivity of the High Plains aquifer re-evaluated using surface drainage patterns. *Geophysical Research Letters*, *39*, L02402. https://doi.org/10.1029/2011GL050200

Maupin, M. A., & Barber, N. L. (2005). *Estimated Withdrawals from Principal Aquifers in the United States, 2000*, *Circular*, (Vol. 1279). Boise, ID: U.S. Geological Survey.

McGuire, V. (2011). *Water-Level Changes in the High Plains Aquifer, Predevelopment to 2009, 2007–08, and 2008–09, and Change in Water in Storage, Predevelopment to 2009*, *Scientific Investigations Report 2011–5089*. Reston, VA: U.S. Geological Survey.

McGuire, V. L. (2017). *Water-Level and Recoverable Water in Storage Changes, High Plains Aquifer, Predevelopment to 2015 and 2013–15*, *Scientific Investigations Report 2011–5089*. Reston, VA: U.S. Geological Survey.

McMahon, P., Plummer, L., Böhlke, J., Shapiro, S., & Hinkle, S. (2011). A comparison of recharge rates in aquifers of the United States based on groundwater-age data. *Hydrogeology Journal*, *19*(4), 779–800. https://doi.org/10.1007/s10040-011-0722-5

Me, W., Abell, J. M., & Hamilton, D. P. (2015). Effects of hydrologic conditions on SWAT model performance and parameter sensitivity for a small, mixed land use catchment in New Zealand. *Hydrology and Earth System Sciences*, *19*(10), 4127–4147. https://doi.org/10.5194/hess-19-4127-2015

Moriasi, D. N., Arnold, J. G., Van Liew, M. W., Bingner, R. L., Harmel, R. D., & Veith, T. L. (2007). Model evaluation guidelines for systematic quantification of accuracy in watershed simulations. *Transactions of the ASABE*, *50*(3), 885–900. https://doi.org/10.13031/2013.23153

National Agricultural Statistics Service (2011). National Agricultural Statistics Service database. Available at http://www.nass.usda.gov/index.asp

Nayak, P. C., Satyaji Rao, Y. R., & Sudheer, K. P. (2006). Groundwater level forecasting in a shallow aquifer using artificial neural network approach. *Water Resources Management*, *20*(1), 77–90. https://doi.org/10.1007/s11269-006-4007-z

Ozdogan, M., & Gutman, G. (2008). A new methodology to map irrigated areas using multi-temporal MODIS and ancillary data: An application example in the continental US. *Remote Sensing of Environment*, *112*(9), 3520–3537. https://doi.org/10.1016/j.rse.2008.04.010

Pianosi, F., Beven, K., Freer, J., Hall, J. W., Rougier, J., Stephenson, D. B., & Wagener, T. (2016). Sensitivity analysis of environmental models: A systematic review with practical workflow. *Environmental Modelling & Software*, *79*, 214–232. https://doi.org/10.1016/j.envsoft.2016.02.008

Portmann, F. T., Siebert, S., & Döll, P. (2010). MIRCA2000—Global monthly irrigated and rainfed crop areas around the year 2000: A new high-resolution data set for agricultural and hydrological modeling. *Global Biogeochemical Cycles*, *24*, GB1011. https://doi.org/10.1029/2008GB003435

Rakhshandehroo, G. R., Akbari, H., Afshari Igder, M., & Ostadzadeh, E. (2018). Long-term groundwater-level forecasting in shallow and deep wells using wavelet neural networks trained by an improved harmony search algorithm. *Journal of Hydrologic Engineering*, *23*(2), 04017058. https://doi.org/10.1061/(ASCE)HE.1943-5584.0001591

Rodell, M., Houser, P. R., Jambor, U., Gottschalck, J., Mitchell, K., Meng, C. J., et al. (2004). The Global Land Data Assimilation System. *Bulletin of the American Meteorological Society*, *85*(3), 381–394. https://doi.org/10.1175/BAMS-85-3-381

Sahoo, S., Russo, T., Elliott, J., & Foster, I. (2017). Machine learning algorithms for modeling groundwater level changes in agricultural regions of the U.S. *Water Resources Research*, *53*, 3878–3895. https://doi.org/10.1002/2016WR019933

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 of the United States of America*, *109*(24), 9320–9325. https://doi.org/10.1073/pnas.1200311109

Shiri, J., & KişI, Ö. (2011). Comparison of genetic programming with neuro-fuzzy systems for predicting short-term water table depth fluctuations. *Computers & Geosciences*, *37*(10), 1692–1701. https://doi.org/10.1016/j.cageo.2010.11.010

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. https://doi.org/10.1623/hysj.48.3.399.45291

Solomatine, D. P., Maskey, M., & Shrestha, D. L. (2008). 'Instance-based learning compared to other data-driven methods in hydrological forecasting'. *Hydrological Processes*, *22*, (275–287). https://doi.org/10.1002/hyp.6592

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. https://doi.org/10.1061/(ASCE)1084-0699(2004)9:6(491)

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. https://doi.org/10.5194/hess-20-1405-2016

Sun, A. Y. (2013). Predicting groundwater level changes using GRACE data. *Water Resources Research*, *49*, 5900–5912. https://doi.org/10.1002/wrcr.20421

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*, 3193–3203. https://doi.org/10.1016/j.jhydrol.2014.10.040

U.S. Geological Survey (2015). National Water Information System. USGS groundwater data for the nation. https://waterdata.usgs.gov/nwis/gw (accessed 21 June 2018).

Wada, Y., van Beek, L. P., van Kempen, C. M., Reckman, J. W., Vasak, S., & Bierkens, M. F. (2010). Global depletion of groundwater resources. *Geophysical Research Letters*, *37*, L20402. https://doi.org/10.1029/2010GL044571

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

Wunsch, A., Liesch, T., & Broda, S. (2018). Forecasting groundwater levels using nonlinear autoregressive networks with exogenous input (NARX). *Journal of Hydrology*, *567*, 743–758. https://doi.org/10.1016/j.jhydrol.2018.01.045

Yadav, B., Ch, S., Mathur, S., & Adamowski, J. (2017). Assessing the suitability of extreme learning machines (ELM) for groundwater level prediction. *Journal of Water and Land Development*, *32*(1), 103–112. https://doi.org/10.1515/jwld-2017-0012