

# Predicting groundwater level changes using GRACE data

Alexander Y. Sun<sup>1</sup>

Received 31 December 2012; revised 7 July 2013; accepted 13 July 2013; published 20 September 2013.

[1] The purpose of this work is to investigate the feasibility of downscaling Gravity Recovery and Climate Experiment (GRACE) satellite data for predicting groundwater level changes and, thus, enhancing current capability for sustainable water resources management. In many parts of the world, water management decisions are traditionally informed by in situ observation networks which, unfortunately, have seen a decline in coverage in recent years. Since its launch, GRACE has provided terrestrial water storage change ( $\Delta TWS$ ) data at global and regional scales. The application of GRACE data for local-scale groundwater resources management has been limited because of uncertainties inherent in GRACE data and difficulties in disaggregating various TWS components. In this work, artificial neural network (ANN) models are developed to predict groundwater level changes directly by using a gridded GRACE product and other publicly available hydrometeorological data sets. As a feasibility study, ensemble ANN models are used to predict monthly and seasonal water level changes for several wells located in different regions across the US. Results indicate that GRACE data play a modest but significant role in the performance of ANN ensembles, especially when the cyclic pattern of groundwater hydrograph is disrupted by extreme climate events, such as the recent Midwest droughts. The statistical downscaling approach taken here may be readily integrated into local water resources planning activities.

**Citation:** Sun, A. Y. (2013), Predicting groundwater level changes using GRACE data, *Water Resour. Res.*, 49, 5900–5912, doi:10.1002/wrcr.20421.

## 1. Introduction

[2] Groundwater is the primary source of drinking and irrigation water supplies in many parts of the world. Shallow groundwater influences 22–32% of global land area [Fan *et al.*, 2013]. In the US, groundwater accounts for 20% of the total water withdrawal, providing more than 90% of water used for rural domestic supplies and 40% used for public supplies [Barber, 2009]. Freshwater aquifers are increasingly stressed by population growth and climate extremes, which are of particular concern for arid and semiarid regions that rely exclusively on groundwater. Worldwide, the pace of groundwater depletion has more than doubled during the last several decades, caused primarily by increasing water demand [Konikow, 2011; Wada *et al.*, 2011, 2012]. Thus, to manage the limited groundwater resources in a way that will ensure equitable, sustainable, and economically prudent decisions, we must enhance our existing capability to monitor and predict water availability. Although in situ monitoring networks provide high-resolution estimates, remotely sensed data constitute the

only source of information for assessing water resources in sparsely monitored regions [Alsdorf *et al.*, 2007; National Research Council, 2008]. For this reason, the Gravity Recovery and Climate Experiment (GRACE) satellite mission has attracted much attention since its launch over a decade ago. GRACE observes temporal variations of Earth's gravitational potential which, after removing atmospheric and oceanic effects, is mostly caused by changes in terrestrial water storage or  $\Delta TWS$  [Tapley *et al.*, 2004].

[3] Remote sensing of groundwater storage changes cannot be done directly using the current technologies [Becker, 2006; Brunner *et al.*, 2007]. In the case of the GRACE satellite, it only tracks  $\Delta TWS$  but not changes of individual hydrologic components (e.g., surface water, soil moisture, and groundwater). By far, the most popular approach for inferring groundwater storage changes from GRACE  $\Delta TWS$  has been to isolate and remove contributions of all other TWS components using either auxiliary information or land surface models. The foci of many previous studies were on comparison between the groundwater storage change obtained by disaggregating GRACE data and that by using in situ data, with the main purpose of validating the accuracy of GRACE data at either regional or continental scales [e.g., Henry *et al.*, 2011; Leblanc *et al.*, 2009; Rodell *et al.*, 2009, 2007; Scanlon *et al.*, 2012a; Swenson *et al.*, 2008; Syed *et al.*, 2008; Yeh *et al.*, 2006]. Practical application of GRACE data for local water resources management, especially nowcasting and forecasting, has been limited. A fundamental issue is related to the low spatial ( $>150,000 \text{ km}^2$ ) and temporal ( $>10$  days) resolution of GRACE [Rodell *et al.*, 2007]. Another compounding issue

Additional supporting information may be found in the online version of this article.

<sup>1</sup>Jackson School of Geosciences, Bureau of Economic Geology, University of Texas at Austin, Austin, Texas, USA.

Corresponding author: A. Y. Sun, Jackson School of Geosciences, Bureau of Economic Geology, University of Texas at Austin, University Station, Austin, TX 78713, USA. (alexsunda@gmail.com)

©2013. American Geophysical Union. All Rights Reserved.  
0043-1397/13/10.1002/wrcr.20421

is the quantification of uncertainties arising from postprocessing and disaggregation GRACE signals.

[4] The raw GRACE data are sets of spherical harmonic coefficients describing monthly variations of Earth's gravity field. Postprocessing needs to be done to remove systematic errors (i.e., destriping) and random errors in higher-degree spherical harmonic coefficients not removed by destriping [Landerer and Swenson, 2012]. The postprocessing operations, however, also cause signal degradation. Disaggregation of GRACE data into contributions of different hydrological components tends to introduce additional uncertainties, as summarized in *Seo et al.* [2006]. Quantification of the uncertainties arising from aggregating (in lateral directions) and disaggregating (in the vertical direction) different TWS components is challenging, especially when large-scale hydrological models are used or when in situ observations are averaged over large areas. In the former case, rigorous uncertainty analyses would require probability distributions of inputs, which are hard to assess in practice. Simple techniques for estimation under uncertainty might be useful. For example, *Sun et al.* [2010] applied a robust least squares approach to estimate aquifer specific yield, which requires only bounds of the  $\Delta$ TWS components.

[5] At the present time, incorporating GRACE data directly into local water management decision making activities, such as groundwater allocation and drought management, is unlikely to be fruitful for aforementioned challenges. But can we use the GRACE product indirectly to inform water management at subgrid scales? This question is equivalent to testing the feasibility of downscaling GRACE data.

[6] Dynamic downscaling techniques are model based [Fowler et al., 2007]. Recently, *Houborg et al.* [2012] assimilated GRACE data into the Catchment Land Surface Model using ensemble Kalman smoother and forcing data from North American and Global Land Data Assimilation Systems Phase 2 (NLDAS-2). NLDAS (1/8° and hourly) consists of several land surface models that cover central North America and is constructed from gauge-based observed precipitation, bias-corrected shortwave radiation, and surface meteorology analysis fields. The results of *Houborg et al.* [2012] show statistically significant improvement in drought prediction skills over many parts of the continental US, highlighting the potential value of GRACE data. *Sun et al.* [2012] imposed GRACE observations as constraints when recalibrating a regional-scale groundwater model such that the model-simulated  $\Delta$ TWS should agree with the GRACE-derived  $\Delta$ TWS. In their case, the regional aquifer model was originally developed for long-term water management and transforming the model for transient simulation at monthly intervals was challenging, requiring significant more spatial and temporal information on groundwater use, recharge rates, and boundary fluxes.

[7] As an alternative to model-based downscaling, statistical downscaling techniques explore empirical relationships between large-scale variables (predictors) and local-scale variables of interest (predictands). The empirical relationship is usually given in the form of a nonparametric model or, transfer function, that is “learned” by

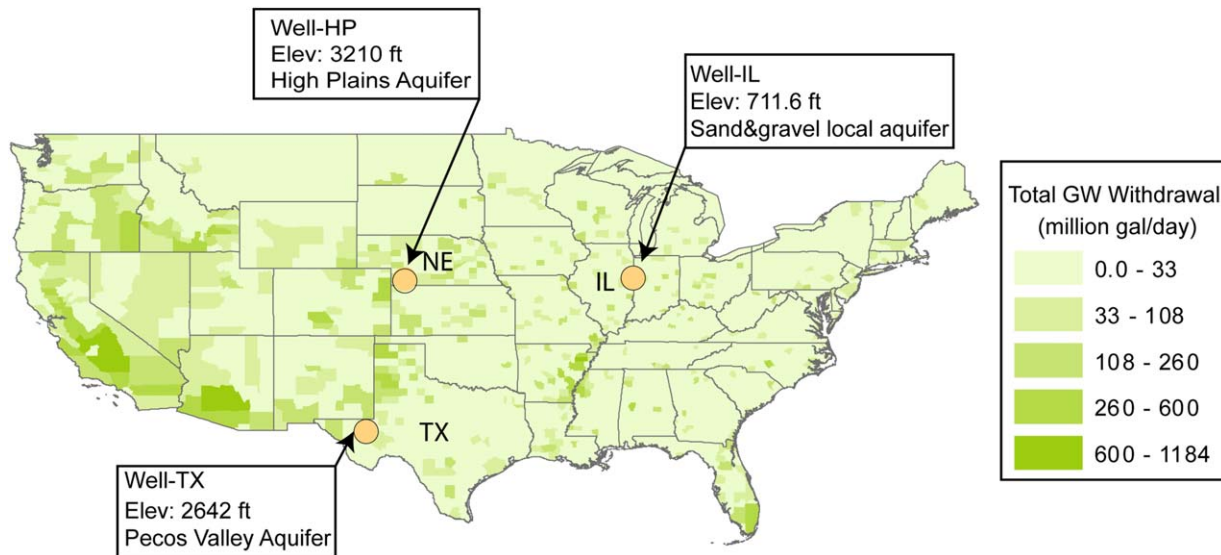
performing linear or nonlinear regression on training data sets. Once trained, the transfer function provides a functional mapping between predictors and predictands that can be used to facilitate cross-scale information exchange. Compared to dynamic downscaling, statistical downscaling is data driven and is much less time consuming to develop [Fowler et al., 2007].

[8] The main hypothesis of this study is that the GRACE  $\Delta$ TWS may serve as a useful predictor for water level changes, in the absence of continuous in situ water level observations. Revolving around this hypothesis, the objectives of this study are to (a) develop statistical downscaling models for incorporating GRACE  $\Delta$ TWS to predict water level changes and (b) validate the usefulness of statistical downscaling, or the “data worth” of GRACE, for wells located in different geographic and climatic regions. Here data worth refers to the contribution of additional observations in reducing predictive uncertainty [Brunner et al., 2012]. In addition to local water resources management needs, this study is also motivated by concerns over the dwindling coverage of in situ monitoring networks, which have been traditionally used to gauge the wellbeing of aquifers, as well as to provide support for local water management decisions (observation wells that support aquifer management decision making are widely known as index wells). The spatiotemporal coverage of in situ monitoring networks depends critically on dispensable local resources. Currently, many water agencies are facing increasing constraints because of budget cuts [Nelson, 2012]. Moreover, even with in situ well networks in operation, holistic groundwater assessments can be difficult to perform on a systematic basis because political and aquifer boundaries typically do not coincide with each other [Rodell et al., 2009]. Thus, insights gained from this analysis by using the freely available GRACE product will be of broad interest to the groundwater community.

[9] The remainder of the paper is organized as follows. Section 2 describes data used and data processing techniques. Section 3 discusses methodologies used for statistical downscaling and forecasting. Section 4 shows results from two sets of downscaling tests. Test I looks at three wells located in different aquifers and climate regions in the U.S. Test II zooms into a single GRACE pixel and further validates the feasibility of GRACE downscaling for multiple sites at subgrid scale. Finally, section 5 summarizes main findings of this study.

## 2. Data and Data Processing

[10] The gridded GRACE product (1° or ~110 km) used in this study is based on Release-5 of GRACE Level-2 data, which has improved accuracy over its precursors due to new data processing algorithms [Landerer and Swenson, 2012]. As mentioned before, filtering and truncation of GRACE spherical harmonic coefficients result in signal attenuation and leakage, which need to be corrected through leakage and bias correction. *Landerer and Swenson* [2012] used scaling to restore signal amplitudes. The spatially distributed scaling field is obtained based on simulation outputs from the Global Land Data Assimilation System (GLDAS). The scaling approach does not seek to match GRACE measurements to GLDAS model



**Figure 1.** Locations of wells considered in the first test. The background map shows the total groundwater withdrawal (in  $10^6$  gals/d) estimated by USGS based on 2005 water use data in the US (Source: National Atlas, <http://www.nationalatlas.gov>).

amplitudes; rather, it uses synthetic model patterns to determine relative signal attenuation based on the ratio of true and filtered signal amplitudes. The availability of this gridded product will enable users who do not possess necessary background in GRACE data processing to extract and use  $\Delta$ TWS time series. The data set is being updated regularly and can be downloaded from Jet Propulsion Laboratory (JPL)'s ftp site [JPL, 2012]. Unless otherwise noted, the period of study spans from March 2003 (the beginning month of the gridded data set) to August 2012 in monthly intervals.

[11] Two tests are designed to validate the concept and significance of GRACE-based downscaling. Test I examines three observation wells located in Texas (Well-TX), Nebraska (Well-HP), and Illinois (Well-IL), respectively (Figure 1). The wells were chosen because (a) all of them have continuous water level records during the study period and (b) at the regional scale, good correlation ( $>0.45$ ) has been found between GRACE-derived groundwater storage changes and those estimated based on in situ water level data for all three aquifers [Scanlon *et al.*, 2012b; Sun *et al.*, 2010; Yeh *et al.*, 2006]. Further background information pertaining to the three wells is provided below.

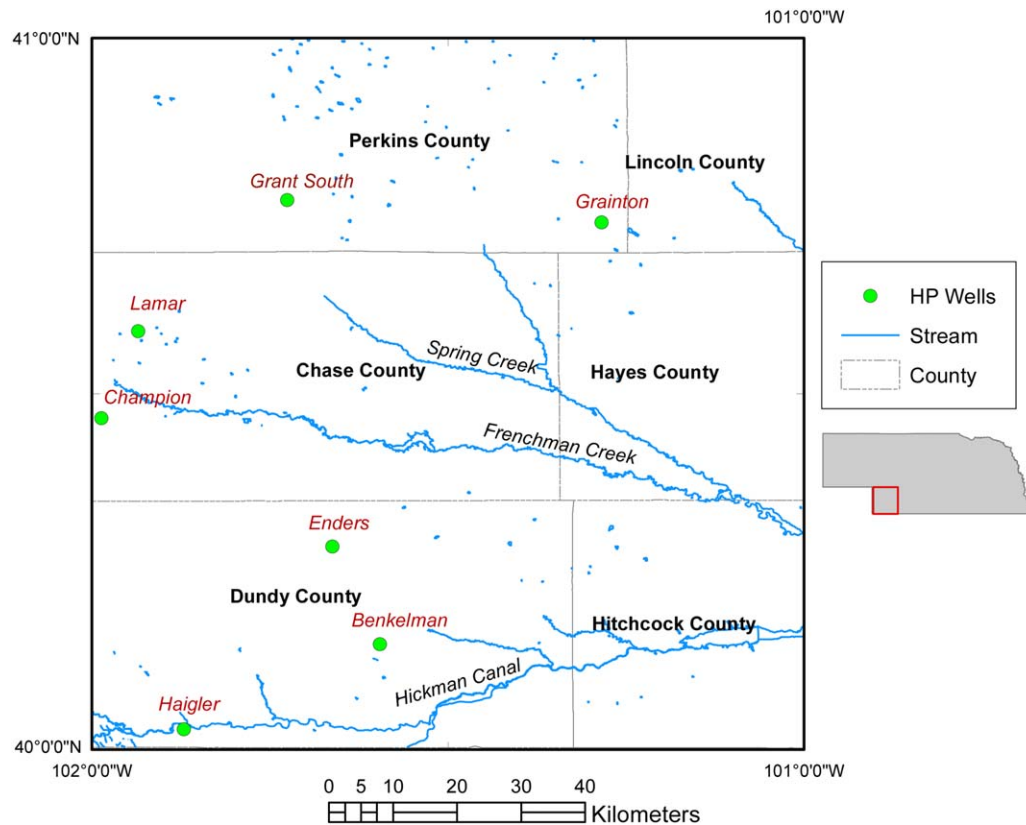
[12] Well-TX is located in Reeves County, Texas, and is completed in the unconfined Pecos Valley alluvial aquifer. The average water level elevation is about 140 ft [ $1 \text{ ft} = 0.3048 \text{ m}$ ] below land surface. According to Texas Water Development Board (TWDB), more than 80% of groundwater pumped from the Pecos Valley aquifer is used for irrigation and the rest is withdrawn for municipal and industrial uses [TWDB, 2012b]. The groundwater level data were queried from TWDB's groundwater database [TWDB, 2012a]. Both Well-HP and Well-IL belong to the US Geologic Survey (USGS)'s Active Groundwater Level network [USGS, 2013]. The former is located in Dundey County, Nebraska, and is completed in the High Plains aquifer, while the latter is located in Vermilion County, Illinois, and is completed in a shallow glacial aquifer. Climate of

the High Plains is semiarid. The unconfined High Plains aquifer is the largest freshwater aquifer in the US. Extensive irrigation in recent decades has led to significant decrease in regional groundwater levels from the predevelopment era (1950s) [Scanlon *et al.*, 2012b; Sophocleous, 2005]. The average water level of Well-HP is 133 ft below land surface during the study period. In comparison, water table of the unconfined Illinois aquifer is relatively shallow and the average water level of Well-IL is only 22 ft below land surface. Rodell and Famiglietti [2001] showed that the groundwater storage change in Illinois is equal in magnitude to soil moisture change, and the two TWS components account for most of the TWS variations in Illinois relative to snow and surface water.

[13] All three Test I wells exhibit a certain degree of seasonality and mild nonstationarity (see supporting information S1, Figure S1). To quantify the correlation between GRACE  $\Delta$ TWS and groundwater level changes, the high-pass, Hodrick-Prescott filter (HP filter) was applied to remove long-term trends from water level data (see section A).

[14] Test II further looks into seven continuously monitored wells that are all located within a  $1^\circ$  GRACE pixel at  $101\text{--}102^\circ\text{W}$  and  $40\text{--}41^\circ\text{N}$  (Figure 2). The area, which is in the Republican River Basin, encompasses several Nebraska counties, including Dundey County where Well-HP (i.e., Benkelman Well in Figure 2) is located. Nebraska is underlain by the largest volume of water-in-storage of the High Plains aquifer [McGuire *et al.*, 2012]. Most of the wells selected for Test II are completed in the Ogallala Formation of the High Plains aquifer. One exception is Haigler Well, which is completed in a relatively shallow sand-and-gravel aquifer. The area is covered by a relatively dense in situ well network and, thus, offers a unique opportunity to further validate the usefulness of GRACE downscaling for multiple sites located within the same GRACE pixel. All well hydrographs were downloaded from USGS web site





**Figure 2.** Map of wells considered in Test II, where all wells are located in the same GRACE pixel in Nebraska. Well names are labeled above the wells.

[USGS, 2013]. More information on Test II well locations and completion depths is provided in Table S1.

[15] In addition to GRACE  $\Delta$ TWS, other hydrometeorological predictors used in this study include the Parameter-elevation Regressions on Independent Slopes Model (PRISM) total monthly precipitation and average minimum and maximum temperature data (4 km resolution) [PRISM Climate Group, 2012], which were extracted for all well locations for the study period.

### 3. Methodology

[16] The statistical downscaling method adopted in this study is artificial neural network (ANN), which has been widely used in streamflow forecasting and water resources management [e.g., ASCE, 2000a, 2000b; Chang and Chang, 2006; House-Peters and Chang, 2011; Hsu et al., 1995; Maier et al., 2010; Moradkhani et al., 2004]. A major strength of the ANN lies in its universal approximation property—an ANN model with a single hidden layer can be trained to approximate the causal relationship of any nonlinear dynamic system without a priori assumptions about the underlying physical system. Another important attribute of ANN is its built-in capability to adapt to changes from the original environment [Haykin, 2009]. Thus, trained ANN models are relatively robust to data noise and are well suited to support real-time decision making.

[17] In the context of water level prediction, Coulibaly et al. [2001] developed ANNs to predict the groundwater

level by using in situ hydrometeorological observations and antecedent groundwater levels from the same well as inputs. Coppola et al. [2003, 2005a] applied ANNs to predict groundwater levels in response to variable pumping and weather conditions. Feng et al. [2008] developed ANNs to simulate regional groundwater levels affected by human activities by incorporating population and monthly irrigation data as additional input variables. Chu and Chang [2009] and Coppola et al. [2007] integrated ANN in an optimization framework for groundwater resources planning, in which ANN models were trained as a surrogate of process models for predicting groundwater levels under different pumping conditions. The models developed in this study differ from previous studies in that the GRACE  $\Delta$ TWS is used in lieu of antecedent groundwater levels as a predictor for 1 month ahead prediction. In other words, the models are developed for situations in which in situ monitoring is disrupted. Because of the GRACE data, the ANN models predict water level changes, instead of absolute water levels. Water agencies are often interested in the former quantity for water availability analysis [McGuire et al., 2012], although estimation of the long-term mean can also be accomplished using a number of existing regression methods [e.g., Hengl et al., 2004] and typically has less requirement on spatiotemporal frequency of data. For completeness, the ANN algorithm is briefly reviewed below. More details on ANN algorithms can be found in Maier et al. [2010] and Haykin [2009].

### 3.1. Multilayer Perceptron (MLP) Networks

[18] Time series prediction problems seek to learn a functional mapping between a set of predictors  $\mathbf{x}$  and the target variable  $y$  (assumed here as a scalar variable)

$$y = f(\mathbf{x}) + \varepsilon, \quad (1)$$

where  $f$  is the mapping and  $\varepsilon$  is process noise. In this work, MLP networks are used to learn the functional mapping. MLP network is a type of feedforward ANN consisting of an input layer, one or more hidden layers, and an output layer, with each layer consisting of one or more neurons. All MLPs developed in this work are single-hidden-layer MLPs. Construction of an MLP proceeds by connecting layers one-by-one through a series of transformations. In the first step, connections between the hidden layer and input variables are established. Let  $\{x_i\}_{i=1}^M$  denote a set of  $M$  predictors. The hidden layer consists of  $K$  hidden neurons and each is a weighted sum of predictors [Bishop, 2006]:

$$a_k = \sum_{i=1}^M w_{ki}^{(1)} x_i + w_{k0}^{(1)}, k = 1, \dots, K, \quad (2)$$

in which  $a_k$  is a hidden neuron,  $\{w_{ki}^{(1)}\}_{i=1}^M$  are unknown weights associated with each input neuron, and  $w_{k0}^{(1)}$  is an unknown bias term used for correcting the estimation bias. The superscripts in equation (2) denote the layer number. In the second step, equation (2) is passed to a transfer function to yield outputs from hidden neurons

$$z_k = \psi(a_k), k = 1, \dots, K, \quad (3)$$

where  $z_k$  are outputs, and  $\psi$  is the transfer function. A commonly used transfer function for the hidden layer is the logistic sigmoid function, which ranges in  $[0, 1]$  and acts as a compressing function. Finally, connections from the hidden layer to the output layer are established via a linear transfer function

$$y_j = \sum_{k=1}^K w_{jk}^{(2)} z_k + w_{j0}^{(2)}, \quad (4)$$

where  $y_j$  ( $j = 1, \dots, J$ ) are output neurons (i.e., model predictions), and  $\{w_{jk}^{(2)}\}_{k=1}^K$  and  $w_{j0}^{(2)}$  are the unknown weights and the bias term of the output layer, respectively. The number of output neurons is always one for the current study. In the training phase, the unknowns in equations (2) and (4) are solved through backpropagation, which is a process of propagating fitting errors backward through the network to obtain the optimal weights in each layer. The Matlab Neural Network Toolbox [Demuth et al., 2008] was used to develop and train all MLP networks for this study.

### 3.2. Performance Measures

[19] Performance of the developed MLPs is quantified using several criteria. Root mean square error (RMSE) measures the global fitness of a predictive model

$$\text{RMSE} = \left( \frac{1}{N_v} \sum_{i=1}^{N_v} (y_i - o_i)^2 \right)^{1/2}, \quad (5)$$

where  $y$  and  $o$  are predicted and observed values, respectively; and  $N_v$  is the number of target data used for testing. The scaled RMSE,  $R^*$ , is defined as the ratio between RMSE and the standard deviation of observations  $\sigma_o$

$$R^* = \frac{\text{RMSE}}{\sigma_o}, \quad (6)$$

which can vary between zero and a large positive value. The correlation coefficient ( $R$ ) is a measure of how future outcomes are likely to be predicted by the model and is equivalent to the sample cross correlation between predicted and observed values

$$R = \frac{\sum_{i=1}^{N_v} (y_i - \bar{y})(o_i - \bar{o})}{\sqrt{\sum_{i=1}^{N_v} (y_i - \bar{y})^2} \sqrt{\sum_{i=1}^{N_v} (o_i - \bar{o})^2}}, \quad (7)$$

where the overbar denotes mean values. The maximum absolute error (MAE) is defined as

$$\text{MAE} = \frac{1}{N_v} \sum_{i=1}^{N_v} |y_i - o_i|. \quad (8)$$

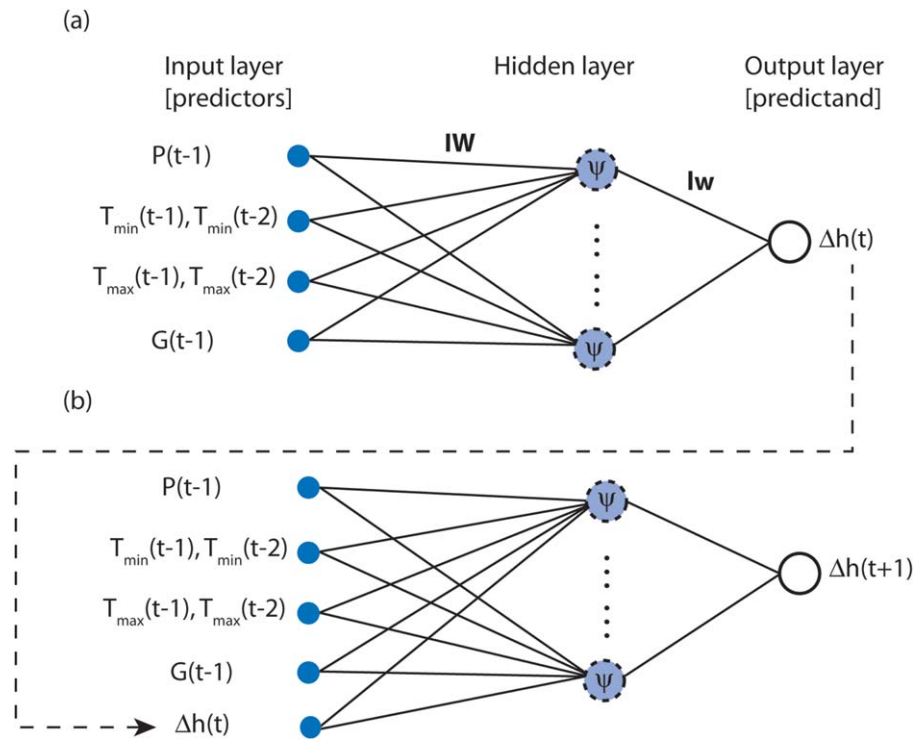
[20] Finally, the Nash-Sutcliffe efficiency (NSE), ranging from  $-\infty$  to 1, measures the predictive skill of a model relative to the mean of observations,

$$\text{NSE} = 1 - \frac{\sum_{i=1}^{N_v} (y_i - o_i)^2}{\sum_{i=1}^{N_v} (o_i - \bar{o})^2}. \quad (9)$$

[21] In the literature, a predictive model is said to achieve (a) *very good* performance if the resulting NSE is greater than 0.75 and  $R^*$  is less than 0.50, (b) *good* performance if NSE is greater than 0.65 and  $R^*$  is less than 0.6, and (c) *satisfactory* performance if NSE is greater than 0.5 and  $R^*$  is less than 0.70 [Moriassi et al., 2007]. These ranking criteria are used here as a convenient measure for comparing the relative performance of different models, although readers should be aware that they are context and application dependent.

### 3.3. ANN Model Structure

[22] Figure 3a shows a representative MLP network structure for 1 month ahead prediction. The input variables  $\mathbf{x}$  includes six variables and the target variable  $y$  is water level change,  $\Delta h$ . Both Well-TX and Well-HP MLPs use the network structure shown in Figure 3a. For Well-IL, an additional antecedent series,  $G(t-2)$ , is included. The MLP structures were determined to maintain a balance between model parsimony and performance. Recall that the main purpose of this work is to test the data worth of



**Figure 3.** Structure of MLP used for (a) 1 month ahead and (b) 2 month ahead predictions. In Figure 3a, the input layer consists of six neurons, which are antecedent precipitation ( $P$ ), minimum and maximum temperature ( $T_{\min}$  and  $T_{\max}$ ), and GRACE  $\Delta TWS$  ( $G$ ). In Figure 3b, the input layer consists of seven neurons. The predictand is water level change ( $\Delta h$ );  $t$  and  $t + 1$  indicate 1 and 2 month ahead;  $IW$  and  $lw$  denote connection weights; and  $\psi$  represents the sigmoid transfer function.

GRACE for fulfilling the role of in situ data. Therefore, the antecedent  $\Delta h$  is not included as a predictor when formulating 1 month ahead prediction MLPs. More details on the design and training of MLPs are provided in section 4.

[23] Water managers are often interested in water level changes at the seasonal scale for operational planning. A strategy proposed by *Chang et al.* [2007] is used here for multistep forecast. The prediction proceeds 1 month at a time, and each time the intermediate output is used to inform the next prediction. This multistep prediction strategy is referred to as serial propagation by *Chang et al.* [2007], who found that inclusion of outputs from intermediate steps not only improved the accuracy, but also reliability of predictions. Thus, output from 1 month ahead prediction is used as a predictor when performing 2 month ahead prediction (Figure 3b). Similarly, both 1 and 2 month ahead  $\Delta h$  predictions are used as inputs for the 3 month ahead prediction. A separate MLP network is developed and trained for each additional prediction step.

### 3.4. Relative Importance of Input Variables

[24] In the literature, several metrics for quantifying the relative importance of ANN input variables exist [*Gevrey et al.*, 2003; *Olden and Jackson*, 2002; *Olden et al.*, 2004], some of which have been applied in previous water level prediction studies. For example, *Coppola et al.* [2003] used the backward stepwise elimination, in which the model performance between excluding and including a certain predictor is compared. However, this classical stepwise method (either forward addition or backward elimination) is limited

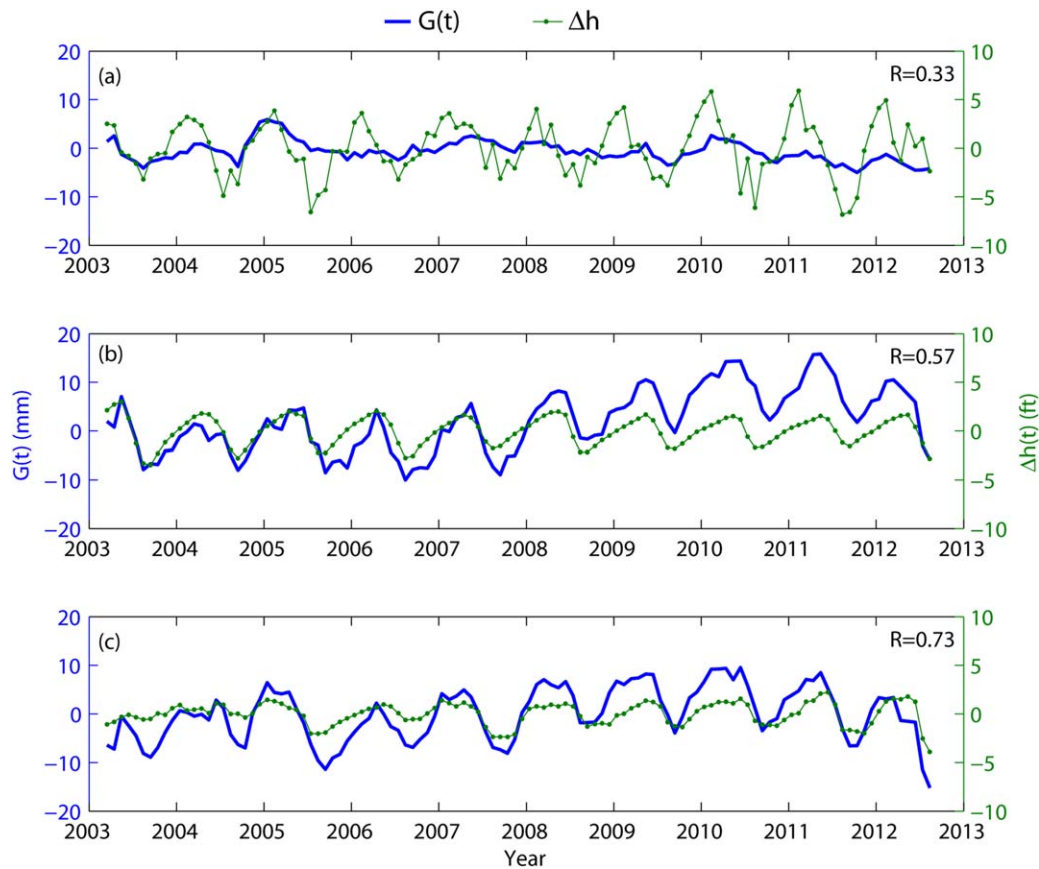
because the “necessity to use a new model for each variable selection skews the results” [*Gevrey et al.*, 2003]. A comparative study of *Olden et al.* [2004] examined the performance of nine relative importance metrics on a synthetic data set with known correlation structure. They concluded that a connection weight method which uses raw input-hidden and hidden-output connection weights in the trained ANN provides the best methodology for accurately quantifying variable importance. The application of the connection weight method proceeds in three steps. Let the input-hidden connection weight matrix be denoted as  $IW_{K \times M}$  and the hidden-output connection weight vector as  $lw_{1 \times K}$ , where  $K$  is the number of hidden neurons, and  $M$  is the number of input variables (see Figure 3a). Each column in  $IW$  is first multiplied by  $lw$  (element-wise multiplication) to give a product matrix  $C_{K \times M}$ . Summing over rows of  $C$  gives an importance vector. The relative importance of the  $i$ th input variable is then defined as follows

$$rel_i = \frac{|v_i|}{\sum_i |v_i|}, i = 1, \dots, K, \quad (10)$$

where  $\{v_i\}_{i=1}^K$  are elements of  $\mathbf{v}$ . The connection weight method is straightforward to implement and will be used here to rank the relative importance of input variables.

### 3.5. Ensemble Generation

[25] Ensemble techniques provide a means for quantifying uncertainty and improving generalization and stability



**Figure 4.** (left axis) GRACE  $\Delta$ TWS and (right axis) groundwater level time series extracted for (a) Well-TX, (b) Well-HP, and (c) Well-IL. Correlation coefficient,  $R$ , is labeled in the upper right corner of each subplot.

of ANN models, which becomes important when the size of training set is small such as in the current problem. Several methods exist for generating ANN ensembles, including bagging [Breiman, 1996], boosting [Freund and Schapire, 1995], and randomization of initial weights [Dietterich, 2000b; Opitz and Maclin, 1999]. Bagging is based on bootstrapping original training sets to generate different realizations; a network model is then trained for each realization. Boosting is a technique for producing a series of predictors trained with a different distribution of the original training data. Initial weight randomization simply uses different initial weights to train and generate an ensemble of networks. The difference in ensemble members reflects uncertainty caused by the limited training sample size and the local nature of ANN training algorithms. Previous studies suggest that the initial weight randomization method can yield results as accurate as the more sophisticated bagging and boosting methods [Opitz and Maclin, 1999; Shu and Burn, 2004]. Thus, it is used to create ensembles in this work.

[26] After the ensemble models are generated, they are combined to generate ensemble outputs. Common methods include simple averaging, weighted averaging, and stacking, a survey of which is provided in Zhou [2012]. Because the current problem involves ensemble models of similar performances (i.e., homogeneous learners), the computationally efficient simple averaging is used, in which all members are assigned equal weights. Regardless of the method used to

construct and combining the ensembles, the generalization ability of a network ensemble is often better than that of a single model [Dietterich, 2000a; Zhou, 2012].

## 4. Results and Discussion

### 4.1. Test I

[27] Correlation analysis shows a positive correlation between GRACE  $\Delta$ TWS ( $G(t)$ ) and water level change ( $\Delta h(t)$ ) for all three wells (Figures 4a–4c), with a Spearman's correlation coefficient of 0.33, 0.57, and 0.73, respectively, for Well-TX, Well-HP, and Well-IL. The correlation is the strongest for Well-IL because of its relatively shallow depth, while intermittent temporal lags between  $G(t)$  and  $\Delta h(t)$  can be observed in the case of the other two wells, especially Well-TX. The lags largely result from different water table depths, the vadose zone soil structure, and shallow aquifer geology which, in turn, all affect recharge patterns. In the case of TX,  $G(t)$  shows no obvious seasonal pattern, while distinct seasonal patterns can be observed in the other two wells. Precipitation and temperature, which have been included in many previous studies as predictors [e.g., Coppola et al., 2005b; Coulibaly et al., 2001], exhibit strong seasonal patterns as expected (Figure S2). Results of autocorrelation analysis on predictors, which is part of the ANN model selection process, are provided in supporting information S3 and Figure S3.



**Table 1.** Test I Ensemble Performance Metrics

Lead Month	Architecture <sup>a</sup>	R <sub>train</sub>	R <sub>val</sub>	R <sub>test</sub>	R*	NSE	MAE (ft)
<i>Well-TX</i>							
1	6-3-1	0.88	0.85	0.83	0.55	0.68	1.61
2	7-4-1	0.89	0.86	0.81	0.66	0.55	1.91
3	8-4-1	0.91	0.85	0.80	0.64	0.57	1.82
<i>Well-HP</i>							
1	6-3-1	0.94	0.94	0.92	0.36	0.86	0.37
2	7-5-1	0.95	0.94	0.91	0.45	0.78	0.47
3	8-4-1	0.91	0.89	0.87	0.60	0.62	0.60
<i>Well-IL</i>							
1	7-6-1	0.86	0.76	0.80	0.58	0.65	0.82
2	8-6-1	0.90	0.69	0.77	0.60	0.62	0.88
3	9-4-1	0.88	0.73	0.75	0.67	0.54	0.97

<sup>a</sup>The three numbers represent the number of input, hidden, and output neurons in each layer.

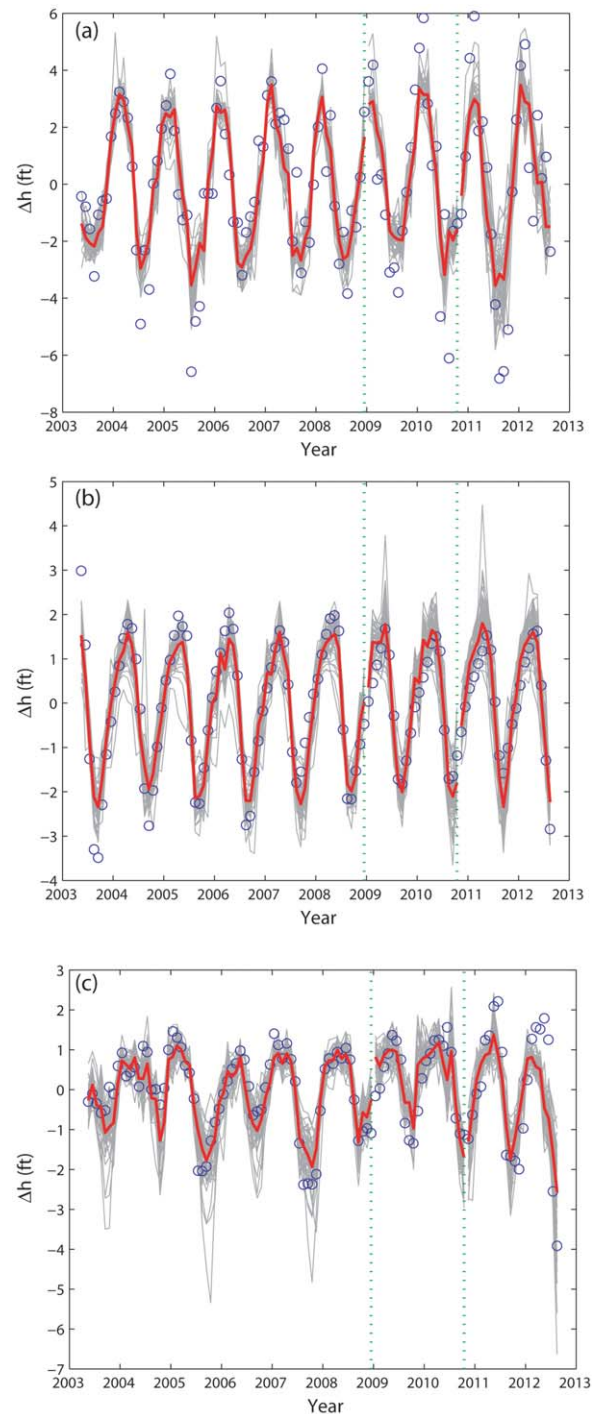
#### 4.1.1. One Month Ahead Prediction

[28] Input data were split into three sequential parts: 60% was used for training, 20% for validation, and the rest for testing the performance of trained models. Before training, all input and target data were standardized by scaling linearly to the range [0.01, 0.99] using the maximum and minimum values of each series. Scaling is necessary to ensure that all variables receive equal consideration during training of the ANN. The number of hidden neurons is critical to the performance of ANN. If too many hidden neurons are used, the network may overfit and, thereby, reducing its generalization capability and increasing training time [Xu and Li, 2002]. On the other hand, using too few hidden neurons may result in underfitting. A rule-of-thumb is that the number of hidden neurons should be about half the number of predictors and should never be more than twice as large [Berry and Linoff, 2004; Minns and Hall, 1996]. In practice, the optimal number is determined by trial-and-error, in which hidden neurons are added gradually until the RMSE of training starts to increase. Following this general procedure, the number of hidden neurons used for Well-TX, Well-HP, and Well-IL models was found to be 3, 3, and 6, respectively. The algorithm chosen for backpropagation is Levenberg-Marquardt, a gradient-based algorithm. Training was stopped when either the performance goal was reached or when the error on validation subset increased consecutively for six iterations. Early stopping is a mechanism for preventing overfitting [Demuth et al., 2008].

[29] For each well and each lead time, an ensemble of 50 MLPs were generated and combined according to the procedure described in section 3.5. The performance metrics of the ensemble models are reported in Table 1, in which the right three columns were calculated based on the testing data. On the basis of NSE, the performance of 1 month ahead ensemble falls into either *good* or *very good* categories (see section 3.2 for definition of performance categories). The fit for Well-HP is the best and that for Well-IL is the worst. Same conclusion can be drawn on the basis of R\*. Figures 5a–5c compare ensemble prediction of  $\Delta h$  to in situ data. On each subplot, the two vertical lines mark the division of training, validation, and testing subsets.

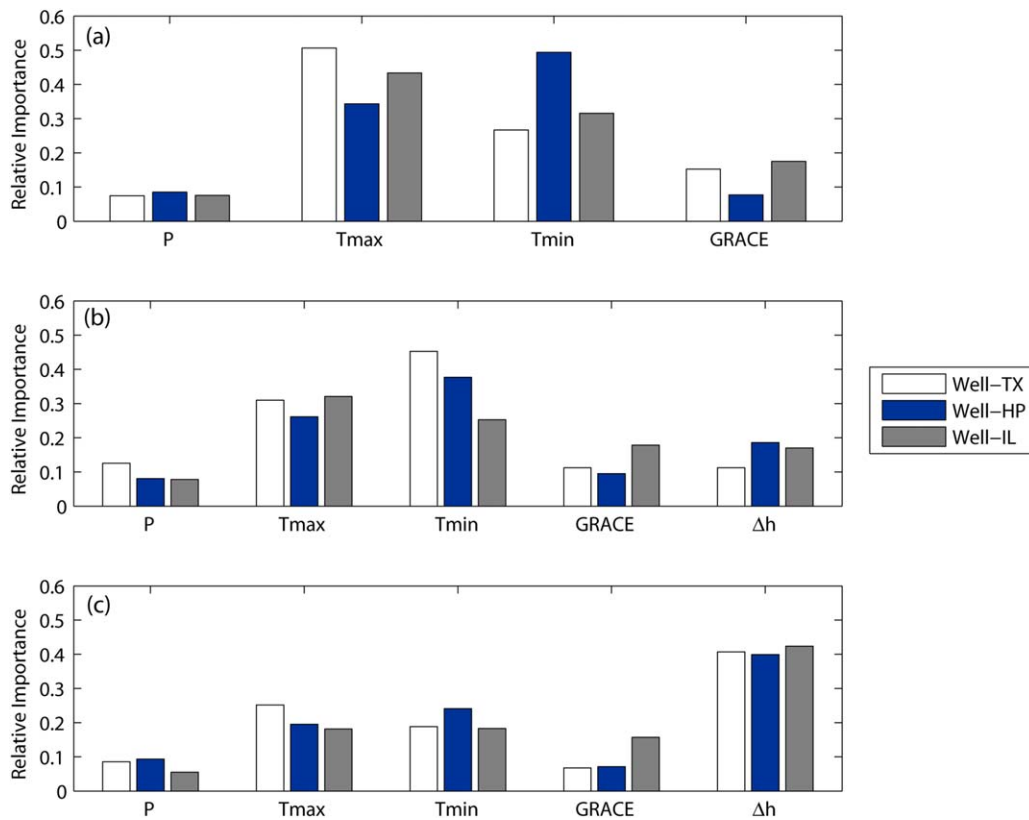
[30] The relatively poor performance of the Well-IL model is surprising because Well-IL shows the strongest  $\Delta h$  versus  $G$  correlation among the three wells. The poor performance may be contributed to both a sharp decline in

water levels during the testing period and a change in seasonal variation patterns from a subdued (2003–2008) to a more pronounced one (2008–2012) (see Figure 4c). In the case of Well-TX, the 2006 drought event was covered by the training period, which helped model training to some



**Figure 5.** One step ahead  $\Delta h$  prediction for (a) Well-TX, (b) Well-HP, and (c) Well-IL. The dotted vertical lines denote boundaries among training, validation, and test subsets. Gray hairlines are predictions of 50 ensemble members, ensemble averages are thick solid lines, and data are labeled by circles.





**Figure 6.** Relative importance of input variables in (a) 1 month, (b) 2 month, and (c) 3 month ahead prediction for Test I, where  $\Delta h$  represents water level change from intermediate steps.

degree, although the developed models need to be further improved to fully capture the extreme values. For the latter purpose, it is necessary to include as many extreme events as training patterns if the corresponding in situ data are also available. It is also important to realize that GRACE monitors the average large-scale impact caused by natural and anthropogenic events. Thus, if an observation well is significantly affected by local pumping activities, GRACE data may have limited value for it. In contrast to Well-TX and Well-IL, Well-HP shows persistent cyclic patterns throughout the study period, which makes its  $\Delta h$  the easiest to predict.

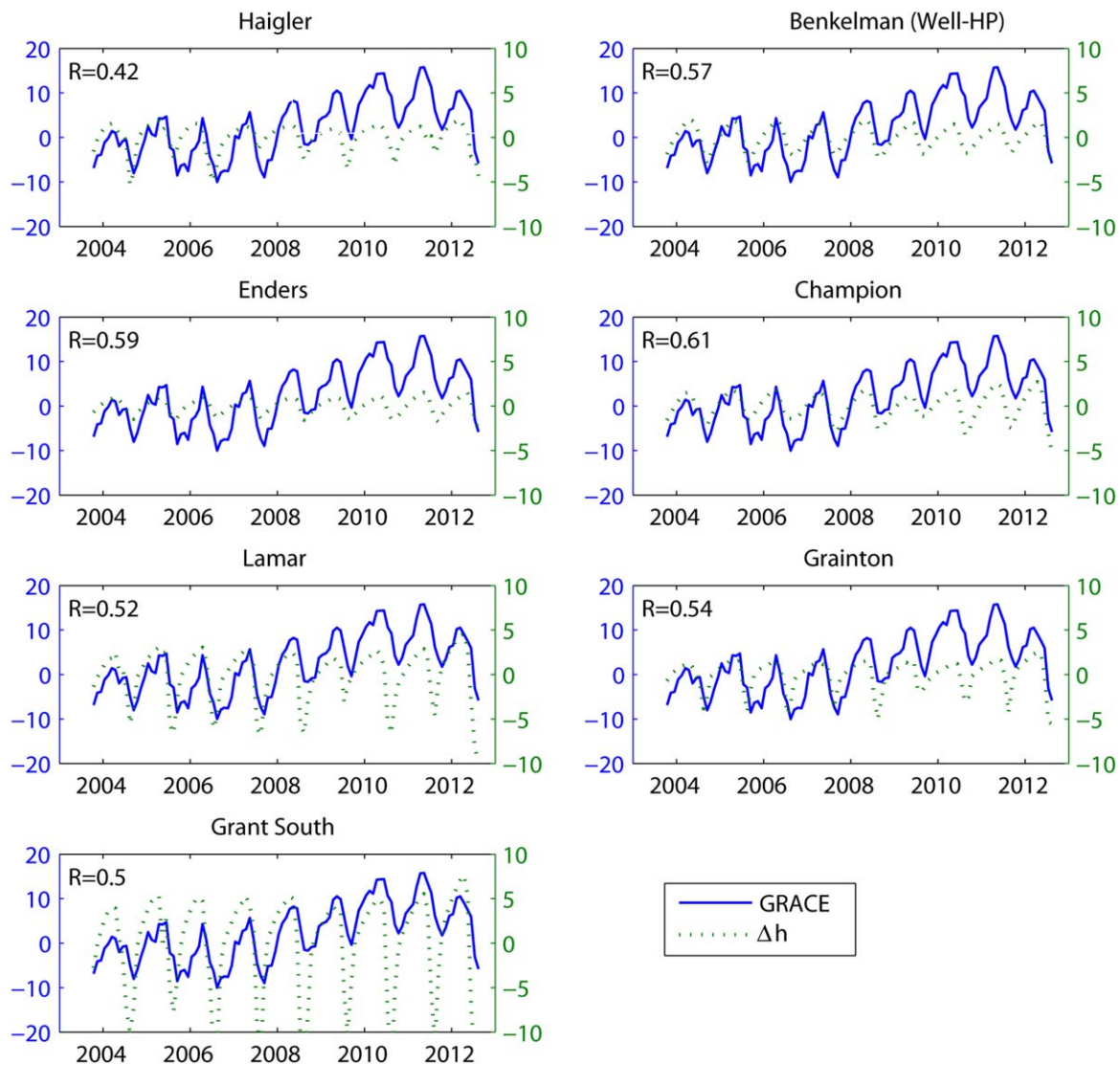
[31] The relative importance of each predictor was quantified using the connection weight method described under section 3.4 and the results are shown in Figure 6a. For ease of visualization, relative importance belonging to the same variable group (e.g.,  $T_{\max}(t-1)$  and  $T_{\max}(t-2)$ ) is combined. The relative importance of  $G$  in 1 month ahead prediction is 0.15, 0.08, and 0.18, respectively, for Well-TX, Well-HP, and Well-IL. In all cases, the contribution of GRACE is similar to or slightly greater than that of the precipitation group, but both are dominated by the temperature group which represents the most important predictor group for 1 month ahead prediction. The relative importance of GRACE appears to be higher when the  $\Delta h$  pattern is less cyclic, such as in the case of Well-TX and Well-IL. Temperature is a good surrogate for seasonality and, thus, it also reflects potential irrigation water uses. The roles of temperature and precipitation as observed here are consist-

ent with those reported elsewhere in previous studies [Coppola et al., 2005b; Feng et al., 2008].

#### 4.1.2. Multimonth Ahead Prediction

[32] Two and three month ahead MLP models were developed according to descriptions in section 3.3 and Figure 3. The ensemble performance metrics are summarized in Table 1. As mentioned before, the intermediate water level change prediction,  $\Delta h(t)$ , is used as feedback in 2 month ahead prediction (i.e., at  $t+1$ ), and intermediate predictions  $\Delta h(t)$  and  $\Delta h(t+1)$  are used for 3 month prediction (i.e., at  $t+2$ ).

[33] The performance of the 2 month ahead prediction spans from *satisfactory* to *very good* category, while the performance of the 3 month ahead predictions all falls into the *satisfactory* category. Figures 6b and 6c plot the relative importance of input variables. For the 2 month ahead prediction, the relative importance of GRACE data stays almost the same as that in the 1 month ahead prediction. The relative importance of intermediate output,  $\Delta h(t)$ , is second only to the temperature group. For the 3 month ahead prediction, the intermediate water level change predictions become the most important group. This is because of the strong autocorrelation of water level [Coppola et al., 2005a]. Thus, in the absence of in situ  $\Delta h$  data, the role of GRACE in 1 month ahead prediction is seen as helping to jumpstart a serial propagation process. The outputs then provide feedbacks to subsequent multimonth ahead prediction. This way, the cross-scale information exchange is facilitated.



**Figure 7.** Comparison between GRACE ( $G$ ) and  $\Delta h$  for all Test II wells. Correlation coefficients are labeled in each subplot.

[34] Overall, Test I results are encouraging. Downscaling of GRACE data leads to modest but significant improvement in  $\Delta h$ . Only one representative well in each different GRACE pixel was selected for testing. This leads naturally to the question about the value of downscaling GRACE for multiple wells at the subgrid level. Test II is designed to address this question.

**Table 2.** One Month Ahead Performance Metrics for Test II Wells

Well Name	$R^*$	NSE	MAE (ft)
Haigler	0.51	0.73	0.72
Benkelman	0.36	0.86	0.37
Enders	0.41	0.83	0.39
Champion	0.51	0.73	0.84
Lamar	0.56	0.67	1.82
Grinton	0.48	0.76	0.95
Grant South	0.44	0.80	2.42

#### 4.2. Test II

[35] The 1° GRACE pixel used for Test II was selected mainly because of the abundance of in situ data that can be used for validation. As can be seen from Figure 2, the wells are relatively distributed across the pixel. Correlation analysis shows that a positive correlation between  $G$  and  $\Delta h$  exists for all seven wells in Test II (Figure 7), with Champion Well showing the strongest (0.61) and Haigler Well the weakest (0.42) correlation. Recall that Haigler Well is completed in a different formation than other wells. Nevertheless, the phases of  $\Delta h$  for all wells are quite similar, although the magnitudes of variations differ among wells. Grant South Well exhibits the largest variations, while variations in other wells are more subdued. For testing purpose, only 1 month ahead MLPs were developed, which adopt the same model structure as that used for Well-HP (i.e., Benkelman Well). Figure 7 also suggests that GRACE data show a slightly upward trend during 2008–2011, which corresponds to a slowdown in groundwater depletion

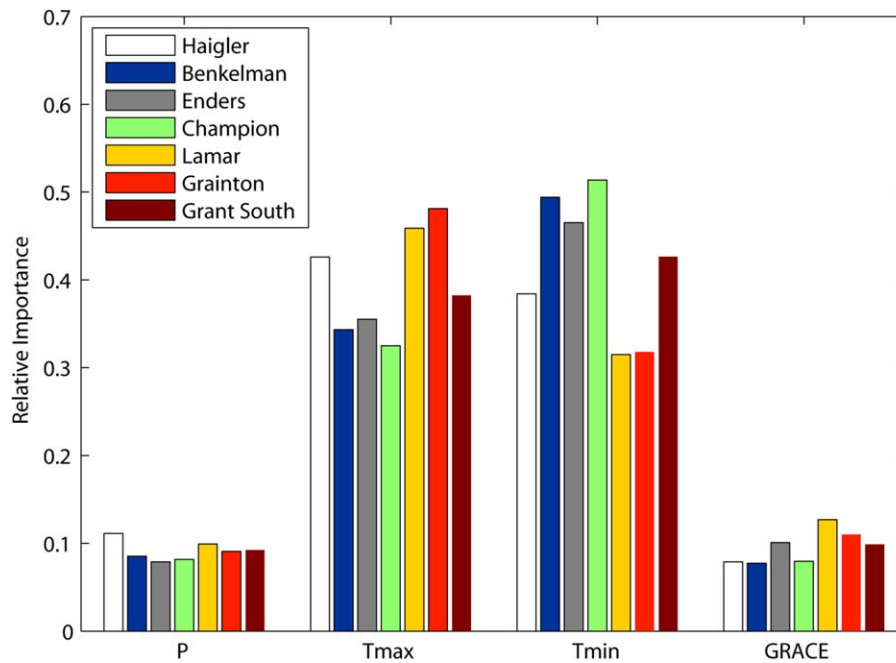


Figure 8. Relative importance of predictors used in Test II.

(Figure S1). The performance metrics of all ensemble models are reported in Table 2.

[36] The NSE values of all models are in either *good* or *very good* category. Benkelman and Enders Well are relatively close, and the ANN models of which show similar performance. Lamar Well and Champion Well, which are located in higher elevations, have lower NSE than others. Figure 8 shows the relative importance of each predictor group in the 1 month ahead MLPs. Again, GRACE has similar relative importance as the precipitation group, and both are dominated by the temperature groups.

[37] Thus, Test II indicates that downscaling of GRACE to multiple wells at the subgrid scale is feasible and has certain merits in this case. This is mainly because most of the wells are completed in the same formation. Of course, if wells are completed in different and disconnected formations, downscaling may have mixed results. Only one pixel was tested here. If data are available, further validation tests for other regions around the world can yield additional insights about the usefulness of GRACE downscaling.

## 5. Conclusion

[38] GRACE has supported many advances in TWS monitoring since its launch. To realize the full potential of GRACE for hydrological applications, however, monthly TWS anomalies from GRACE must be downscaled in space and time and extrapolated to the present, thereby meeting the specificity, timeliness, and high spatial resolution requirements of local water resources management [Houborg *et al.*, 2012]. This study provides a relatively straightforward nonparametric approach for tapping into the predictive power of GRACE.

[39] A newly released, gridded GRACE product was tested for its value in serving as a surrogate for continuously monitored in situ water level measurements, which

have seen a great decline in coverage decline. Groundwater agencies, especially those in semiarid regions, are interested in tracking groundwater level changes to manage aquifers in a sustainable manner [McGuire *et al.*, 2012]. The nonparametric ANN models used in this study were developed using PRISM monthly precipitation, maximum and minimum temperatures, and GRACE  $\Delta$ TWS as inputs. The target variable of interest is the groundwater level change,  $\Delta h$ . The wells tested are located in different geographic and climatic regions in the US. Main findings include:

[40] 1. At 1 month prediction interval, the trained ensemble MLPs gave *good* to *very good* performance. The relative importance of GRACE to  $\Delta h$  prediction ranges from 8% to nearly 20%. The relative importance of GRACE tends to be greater when the water level patterns are less cyclic.

[41] 2. The performance of the ensemble models is satisfactory at multimonth lead times.

[42] 3. Results support the main hypothesis that GRACE  $\Delta$ TWS can be downscaled to infer or predict  $\Delta h$  when continuous in situ measurements are not available.

[43] 4. The approach developed here can be applied to force multiple ANNs developed for a network of wells, the outputs of which can then be combined via spatial interpolation techniques.

[44] 5. The ANN approach taken here is well established, although its training and validation can be time consumption. However, any other statistical learning method can be used in its place if necessary.

## Appendix A

[45] The Hodrick–Prescott filter [Hodrick and Prescott, 1997] or HP filter is used to obtain  $\Delta h$  from absolute water level measurements. HP filter is a high-pass filter that removes a smooth trend from a time series by solving [Ravn and Uhlig, 2002]



$$\min \sum_{t=1}^T \left\{ (y_t - \bar{y}_t)^2 + \lambda [(\bar{y}_{t+1} - \bar{y}_t) - (\bar{y}_t - \bar{y}_{t-1})]^2 \right\}, \quad (\text{A1})$$

where  $\bar{y}_t$  is the trend term and  $\lambda$  is a smoothing parameter. If the frequency of raw data is monthly, the recommended periodicity parameter of the HP filter is 14,400 [Ravn and Uhlig, 2002]. The Matlab function `hpfiler` was used in this study.

[46] **Acknowledgments.** The author is grateful to the Associate Editor and three anonymous reviewers for their careful review and constructive comments, which have significantly improved the original manuscript. The author also wishes to thank Sean Swenson at the National Center for Atmospheric Research for his constructive comments on the original manuscript.

## References

- Alsodorf, D. E., E. Rodríguez, and D. P. Lettenmaier (2007), Measuring surface water from space, *Rev. Geophys.*, 45, RG2002, doi:10.1029/2006RG000197.
- ASCE (2000a), Committee on the application of ANNs in hydrology artificial neural networks in hydrology. I: Preliminary concepts, *J. Hydrol. Eng.*, 5(2), 115–123.
- ASCE (2000b), Committee on the application of ANNs in hydrology artificial neural networks in hydrology. II: Hydrologic application, *J. Hydrol. Eng.*, 5(2), 124–137.
- Barber, N. L. (2009), Summary of estimated water use in the United States in 2005: U.S. Geological Survey Fact Sheet 2009–3098, 2 pp., U.S. Geol. Surv., Reston, Va.
- Becker, M. W. (2006), Potential for satellite remote sensing of ground water, *Ground Water*, 44(2), 306–318.
- Berry, M. J., and G. S. Linoff (2004), *Data Mining Techniques: For Marketing, Sales, and Customer Relationship Management*, John Wiley, Indianapolis, IN.
- Bishop, C. M. (2006), *Pattern Recognition and Machine Learning*, 738 p., Springer, New York.
- Breiman, L. (1996), Bagging predictors, *Mach. Learning*, 24(2), 123–140.
- Brunner, P., H.-J. H. Franssen, L. Kgotlhang, P. Bauer-Gottwein, and W. Kinzelbach (2007), How can remote sensing contribute in groundwater modeling?, *Hydrogeol. J.*, 15(1), 5–18.
- Brunner, P., J. Doherty, and C. T. Simmons (2012), Uncertainty assessment and implications for data acquisition in support of integrated hydrologic models, *Water Resour. Res.*, 48, W07513, doi:10.1029/2011WR011342.
- Chang, F. J., and Y. T. Chang (2006), Adaptive neuro-fuzzy inference system for prediction of water level in reservoir, *Adv. Water Resour.*, 29(1), 1–10.
- Chang, F. J., Y.-M. Chiang, and L.-C. Chang (2007), Multi-step-ahead neural networks for flood forecasting, *Hydrol. Sci. J.*, 52(1), 114–130.
- Chu, H. J., and L. C. Chang (2009), Optimal control algorithm and neural network for dynamic groundwater management, *Hydrol. Processes*, 23(19), 2765–2773.
- Coppola, E. A., F. Szidarovszky, M. Poulton, and E. Charles (2003), Artificial neural network approach for predicting transient water levels in a multilayered groundwater system under variable state, pumping, and climate conditions, *J. Hydrol. Eng.*, 8(6), 348–360.
- Coppola, E. A., A. J. Rana, M. M. Poulton, F. Szidarovszky, and V. W. Uhl (2005a), A neural network model for predicting aquifer water level elevations, *Ground Water*, 43(2), 231–241.
- Coppola, E. A., C. F. McLane, M. M. Poulton, F. Szidarovszky, and R. D. Magelky (2005b), Predicting conductance due to upconing using neural networks, *Ground Water*, 43(6), 827–836.
- Coppola, E. A., F. Szidarovszky, D. Davis, S. Spayd, M. M. Poulton, and E. Roman (2007), Multiobjective analysis of a public wellfield using artificial neural networks, *Ground Water*, 45(1), 53–61.
- Coulibaly, P., F. Anctil, R. Aravena, and B. Bobee (2001), Artificial neural network modeling of water table depth fluctuations, *Water Resour. Res.*, 37(4), 885–896.
- Demuth, H., M. Beale, and M. Hagan (2008), *Neural Network Toolbox™ User's Guide*, 907 pp., The MathWorks Inc., Natick, Mass.
- Dietterich, T. G. (2000a), *Ensemble methods in machine learning*, in Multiple Classifier Systems, edited Kittler J. and F. Roli, pp. 1–15, Springer, Heidelberg.
- Dietterich, T. G. (2000b), An experimental comparison of three methods for constructing ensembles of decision trees: Bagging, boosting, and randomization, *Mach. Learning*, 40(2), 139–157.
- Fan, Y., H. Li, and G. Miguez-Macho (2013), Global patterns of groundwater table depth, *Science*, 339(6122), 940–943.
- Feng, S., S. Kang, Z. Huo, S. Chen, and X. Mao (2008), Neural networks to simulate regional ground water levels affected by human activities, *Ground Water*, 46(1), 80–90.
- Fowler, H., S. Blenkinsop, and C. Tebaldi (2007), Linking climate change modelling to impacts studies: Recent advances in downscaling techniques for hydrological modelling, *Int. J. Climatol.*, 27(12), 1547–1578.
- Freund, Y., and R. Schapire (1995), A decision-theoretic generalization of on-line learning and an application to boosting, in *Computational Learning Theory*, vol. 904, pp. 23–37, Springer, Berlin.
- Gevrey, M., I. Dimopoulos, and S. Lek (2003), Review and comparison of methods to study the contribution of variables in artificial neural network models, *Ecol. Modell.*, 160(3), 249–264.
- Haykin, S. S. (2009), *Neural Networks and Learning Machines*, 3rd ed., 906 p., Prentice Hall, New York.
- Hengl, T., G. B. M. Heuvelink, and A. Stein (2004), A generic framework for spatial prediction of soil variables based on regression-kriging, *Geoderma*, 120(1–2), 75–93.
- Henry, C. M., D. M. Allen, and J. Huang (2011), Groundwater storage variability and annual recharge using well-hydrograph and GRACE satellite data, *Hydrogeol. J.*, 19(4), 741–755.
- Hodrick, R. J., and E. C. Prescott (1997), Postwar US business cycles: An empirical investigation, *J. Money Credit Banking*, 28, 1–16.
- Houborg, R., M. Rodell, B. Li, R. Reichle, and B. F. Zaitchik (2012), Drought indicators based on model-assimilated Gravity Recovery and Climate Experiment (GRACE) terrestrial water storage observations, *Water Resour. Res.*, 48, W07525, doi:10.1029/2011WR011291.
- House-Peters, L. A., and H. Chang (2011), Urban water demand modeling: Review of concepts, methods, and organizing principles, *Water Resour. Res.*, 47, W05401, doi:10.1029/2010WR009624.
- Hsu, K. L., H. V. Gupta, and S. Sorooshian (1995), Artificial neural-network modeling of the rainfall-runoff process, *Water Resour. Res.*, 31(10), 2517–2530.
- JPL (2012), GRACE product, Pasadena, CA [Available at [http://podaac-ftp.jpl.nasa.gov/allData/tellus/L3/land\\_mass/RL05/](http://podaac-ftp.jpl.nasa.gov/allData/tellus/L3/land_mass/RL05/)].
- Konikow, L. F. (2011), Contribution of global groundwater depletion since 1900 to sea-level rise, *Geophys. Res. Lett.*, 38, L17401, doi:10.1029/2011GL048604.
- Landerer, F. W., and S. C. Swenson (2012), Accuracy of scaled GRACE terrestrial water storage estimates, *Water Resour. Res.*, 48, W04531, doi:10.1029/2011WR011453.
- Leblanc, M. J., P. Tregoning, G. Ramillien, S. O. Tweed, and A. Fakes (2009), Basin-scale, integrated observations of the early 21st century multiyear drought in southeast Australia, *Water Resour. Res.*, 45, W04408, doi:10.1029/2008WR007333.
- Maier, H. R., A. Jain, G. C. Dandy, and K. P. Sudheer (2010), Methods used for the development of neural networks for the prediction of water resource variables in river systems: Current status and future directions, *Environ. Modell. Software*, 25(8), 891–909.
- McGuire, V. L., K. D. Lund, and B. K. Densmore (2012), Saturated thickness and water in storage in the High Plains aquifer, 2009, and water-level changes and changes in water in storage in the High Plains aquifer, 1980 to 1995, 1995 to 2000, 2000 to 2005, and 2005 to 2009, *USGS Sci. Invest. Rep. 2012–5177*, 28 pp., Reston, Va.
- Minns, A., and M. Hall (1996), Artificial neural networks as rainfall-runoff models, *Hydrol. Sci. J.*, 41(3), 399–417.
- Moradkhani, H., K.-L. Hsu, H. V. Gupta, and S. Sorooshian (2004), Improved streamflow forecasting using self-organizing radial basis function artificial neural networks, *J. Hydrol.*, 295(1–4), 246–262.
- Moriasi, D. N., J. G. Arnold, M. W. VanLiew, R. L. Bingner, R. D. Harmel, and T. L. Veith (2007), Model evaluation guidelines for systematic quantification of accuracy in watershed simulations, *Trans. Am. Soc. Agric. Biol. Eng.*, 50(3), 885–900.
- National Research Council (2008), *Integrating Multiscale Observations of U.S. Waters*, 210 pp. The National Academies Press, Washington, D. C.
- Nelson, R. L. (2012), Assessing local planning to control groundwater depletion: California as a microcosm of global issues, *Water Resour. Res.*, 48, W01502, doi:10.1029/2011WR010927.
- Olden, J. D., and D. A. Jackson (2002), Illuminating the “black box”: A randomization approach for understanding variable contributions in artificial neural networks, *Ecol. Modell.*, 154, 135–150.

- Olden, J. D., M. K. Joy, and R. G. Death (2004), An accurate comparison of methods for quantifying variable importance in artificial neural networks using simulated data, *Ecol. Modell.*, 178(3–4), 389–397.
- Opitz, D., and R. Maclin (1999), Popular ensemble methods: An empirical study, *J. Artif. Intel. Res.*, 11, 169–198.
- PRISM Climate Group (2012), PRISM, Corvallis, OR [Available at <http://prism.oregonstate.edu>.]
- Ravn, M. O., and H. Uhlig (2002), On adjusting the Hodrick-Prescott filter for the frequency of observations, *Rev. Econ. Stat.*, 84(2), 371–376.
- Rodell, M., and J. S. Famiglietti (2001), Terrestrial water storage variations over Illinois: Analysis of observations and implications for Gravity Recovery and Climate Experiment (GRACE), *Water Resour. Res.*, 37(5), 1327–1340.
- Rodell, M., J. L. Chen, H. Kato, J. S. Famiglietti, J. Nigro, and C. R. Wilson (2007), Estimating groundwater storage changes in the Mississippi River basin (USA) using GRACE, *Hydrogeol. J.*, 15(1), 159–166.
- Rodell, M., I. Velicogna, and J. S. Famiglietti (2009), Satellite-based estimates of groundwater depletion in India, *Nature*, 460(7258), 999–1002.
- Scanlon, B. R., L. Longuevergne, and D. Long (2012a), Ground referencing GRACE satellite estimates of groundwater storage changes in the California Central Valley, USA, *Water Resour. Res.*, 48, W04520, doi:10.1029/2011WR011312.
- Scanlon, B. R., C. C. Faunt, L. Longuevergne, R. C. Reedy, W. M. Alley, V. L. McGuire, and P. B. McMahon (2012b), Groundwater depletion and sustainability of irrigation in the US High Plains and Central Valley, *Proc. Natl. Acad. Sci.*, 109(24), 9320–9325.
- Seo, K., C. Wilson, J. Famiglietti, J. Chen, and M. Rodell (2006), Terrestrial water mass load changes from Gravity Recovery and Climate Experiment (GRACE), *Water Resour. Res.*, 42, W05417, doi:10.1029/2005WR004255.
- Shu, C., and D. H. Burn (2004), Artificial neural network ensembles and their application in pooled flood frequency analysis, *Water Resour. Res.*, 40, W09301, doi:10.1029/2003WR002816.
- Sophocleous, M. (2005), Groundwater recharge and sustainability in the High Plains aquifer in Kansas, USA, *Hydrogeol. J.*, 13, 351–365.
- Sun, A. Y., R. Green, M. Rodell, and S. Swenson (2010), Inferring aquifer storage parameters using satellite and in situ measurements: Estimation under uncertainty, *Geophys. Res. Lett.*, 37, L10401, doi:10.1029/2010GL043231.
- Sun, A. Y., R. Green, S. Swenson, and M. Rodell (2012), Toward calibration of regional groundwater models using GRACE data, *J. Hydrol.*, 422–423, 1–9.
- Swenson, S., J. Famiglietti, J. Basara, and J. Wahr (2008), Estimating profile soil moisture and groundwater variations using GRACE and Oklahoma Mesonet soil moisture data, *Water Resour. Res.*, 44, W01413, doi:10.1029/2007WR006057.
- Syed, T. H., J. S. Famiglietti, M. Rodell, J. Chen, and C. R. Wilson (2008), Analysis of terrestrial water storage changes from GRACE and GLDAS, *Water Resour. Res.*, 44, W02433, doi:10.1029/2006WR005779.
- Tapley, B. D., S. Bettadpur, M. Watkins, and C. Reigber (2004), The gravity recovery and climate experiment: Mission overview and early results, *Geophys. Res. Lett.*, 31, L09607, doi:10.1029/2004GL019920.
- TWDB (2012a), Groundwater database, Austin, TX. [Available at <http://www.twdb.state.tx.us/groundwater/data/gwdbprpt.asp>.]
- TWDB (2012b), Pecos valley aquifer. [Available at <http://www.twdb.state.tx.us/groundwater/aquifer/majors/pecos-valley.asp>.]
- USGS (2013), USGS active groundwater level network, Reston, Va. [Available at <http://groundwaterwatch.usgs.gov>.]
- Wada, Y., L. VanBeek, D. Viviroli, H. H. Dürr, R. Weingartner, and M. F. P. Bierkens (2011), Global monthly water stress: 2. Water demand and severity of water stress, *Water Resour. Res.*, 47, W07518, doi:10.1029/2010WR009792.
- Wada, Y., L. P. H. vanBeek, F. C. Sperna Weiland, B. F. Chao, Y.-H. Wu, and M. F. P. Bierkens (2012), Past and future contribution of global groundwater depletion to sea-level rise, *Geophys. Res. Lett.*, 39, L09402, doi:10.1029/2012GL051230.
- Xu, Z., and J. Li (2002), Short-term inflow forecasting using an artificial neural network model, *Hydrol. Processes*, 16(12), 2423–2439.
- Yeh, P. J. F., S. C. Swenson, J. S. Famiglietti, and M. Rodell (2006), Remote sensing of groundwater storage changes in Illinois using the Gravity Recovery and Climate Experiment (GRACE), *Water Resour. Res.*, 42, W12203, doi:10.1029/2006WR005374.
- Zhou, Z.-H. (2012), *Ensemble Methods: Foundations and Algorithms*, vol. 14, 222 p., Taylor and Francis, Boca Raton, Fla.