Groundwater Withdrawal Prediction Using Integrated Multitemporal Remote Sensing Data Sets and Machine Learning

S. Majumdar¹, R. Smith¹, J. J. Butler Jr.², and V. Lakshmi³

¹Geosciences and Geological and Petroleum Engineering (GGPE) Department, Missouri University of Science and Technology, Rolla, MO, USA, ²Kansas Geological Survey, University of Kansas, Lawrence, KS, USA, ³Department of Engineering Systems and the Environment, University of Virginia, Charlottesville, VA, USA

Abstract Effective monitoring of groundwater withdrawals is necessary to help mitigate the negative impacts of aquifer depletion. In this study, we develop a holistic approach that combines water balance components with a machine learning model to estimate groundwater withdrawals. We use both multitemporal satellite and modeled data from sensors that measure different components of the water balance and land use at varying spatial and temporal resolutions. These remote sensing products include evapotranspiration, precipitation, and land cover. Due to the inherent complexity of integrating these data sets and subsequently relating them to groundwater withdrawals using physical models, we apply random forests—a state of the art machine learning algorithm—to overcome such limitations. Here, we predict groundwater withdrawals per unit area over a highly monitored portion of the High Plains aquifer in the central United States at 5 km resolution for the Years 2002–2019. Our modeled withdrawals had high accuracy on both training and testing data sets ($R^2 \approx 0.99$ and $R^2 \approx 0.93$, respectively) during leave-one-out (year) cross validation with low mean absolute error (MAE) $\approx 4.31$ mm and root-mean-square error (RMSE) $\approx 13.50$ mm for the year 2014. Moreover, we found that even for the extreme drought year of 2012, we have a satisfactory test score ($R^2 \approx 0.84$) with MAE $\approx 9.72$ mm and RMSE $\approx 24.17$ mm. Therefore, the proposed machine learning approach should be applicable to similar regions for proactive water management practices.

Plain Language Summary Groundwater is an essential component of global water resources and is the largest source of Earth's liquid freshwater. It is extensively used for drinking water and to support global food production. Consequently, groundwater consumption has significantly increased owing to the pressing demands for water, food, and energy primarily driven by the increasing global population. Despite its critical role in the water-food-energy nexus, very few regions in the United States or elsewhere actively monitor their groundwater withdrawals (extraction or pumping) for implementing sustainable water management solutions. We develop a novel approach combining water balance components and land use measured using openly available remote sensing products with a machine learning model to estimate groundwater withdrawals. This framework automatically learns the interrelationships among these variables and groundwater withdrawals. Our study area is a portion of the High Plains Aquifer in Kansas (central United States) where overpumping has caused substantial groundwater storage loss. Also, a large amount of groundwater pumping data are available for validation. The results indicate good accuracy even for extreme drought years. Thus, this approach should be applicable to similar regions having sparsely or moderately available groundwater pumping data, enabling water managers to proactively implement sustainable solutions addressing water security issues.

1. Introduction

Groundwater constitutes nearly 30% of total global freshwater reserves (Schneider et al., 2011) and is a key component of the water-food-energy nexus (Smajgl et al., 2016). Globally, about half of the drinking water is supplied by groundwater (GW) (Gleeson et al., 2019). Moreover, GW is extensively used in agricultural activities, and the demand for it is increasing due to the rise in global population, dietary shifts, and climate change (Margat & van der Gun, 2013). Monitoring of GW withdrawals (pumping) is needed for quantifying
aquifer depletion and to provide essential information for building GW models to manage the resource. However, most places in the world, including the United States, do not actively monitor their GW withdrawals at a high-enough spatial resolution for implementing sustainable water management policies. Without that monitoring, it will be exceedingly difficult to address the negative impacts of GW overuse, which include permanent aquifer storage loss, land subsidence, and water contamination (Butler et al., 2018; Chaussard & Farr, 2019; Erban et al., 2013; Galloway & Burbey, 2011; MardanDoost et al., 2019; Smith et al., 2017, 2018; Smith & Majumdar, 2020).

Existing methods for estimating GW pumping rely on estimates of agricultural water demand, typically done with evapotranspiration (ET) and soil models, and surface water availability. The Central Valley Hydrologic Model (CVHM) is a sophisticated example of this approach. Faunt (2009) used the FMP package in MODFLOW (Schmid, 2004) to simulate water demand given land use data and an ET model that considers root depth, temperature, precipitation, and crop type. The model also uses known surface water availability and apportions the remaining water demand to GW pumping. Dogrul et al. (2016) also developed a model to simulate water demand of irrigated crops. While these models are beneficial and have been used successfully, they also require a significant amount of parameters and ground-truth data to calibrate, which makes them time-consuming and expensive to apply over large geographic areas.

Due to the availability of voluminous amounts of satellite data, it is possible to monitor large regions using remote sensing techniques (Frappart & Bourrel, 2018; Leidner & Buchanan, 2018). In the hydrologic remote sensing domain, there is a multitude of spaceborne missions that provide satellite products for assessing different quantities related to the global water cycle. The most prominent of these products quantify total water storage (TWS)—GRACE (Gravity Recovery and Climate Experiment) and GRACE-FO (GRACE—Follow On), terrestrial ET—MODIS (Moderate Resolution Imaging Spectroradiometer), soil moisture—SMAP (Soil Moisture Active Passive), precipitation—GPM (Global Precipitation Measurement), TRMM (Tropical Rainfall Measuring Mission), and land use—USDA-NASS (United States Department of Agriculture-National Agricultural Statistics Service) (Boryan et al., 2011; Chan et al., 2016; MardanDoost et al., 2019; Nie et al., 2018; Yi et al., 2018).

Each of these data sets is related in some way to hydrologic fluxes that impact the GW system. A great deal of research has been done on using these data sets to estimate GW fluxes. A number of studies have used GRACE to estimate changes in TWS then subtract the components of soil moisture, surface water, and snow water to estimate fluxes in GW storage (Famiglietti et al., 2011; Rodell et al., 2007, 2009). While useful for basin- or continental-scale studies, the resolution of GRACE is too coarse (~400 km) for local estimates of GW flux. Surface water availability and land use (which is a proxy for water demand) have also been tied to GW withdrawals in previous studies (Faunt, 2009). Many land use data sets are now derived from remote sensing, and a growing number of studies are using remote sensing estimates of land use to estimate irrigated area (Deines et al., 2017; Özdogan & Gutman, 2008), although this has not been tied directly to GW extraction. Satellite methods have been used to estimate GW storage change at high resolution (~100 m) utilizing land subsidence estimates from interferometric synthetic aperture radar (InSAR, e.g., Hoffmann et al., 2001; Smith et al., 2017), but this approach is typically limited to regions with confined or semiconfined aquifers (where pressure changes are highest) and highly compressible sediments (Smith & Majumdar, 2020). In spite of the mature research in many of these individual fields, very rarely are the various remotely sensed data sets combined to estimate GW fluxes, which limits their ability to estimate these fluxes to very specific cases. In this study, we seek a generalizable approach that utilizes satellite data to estimate GW withdrawals at the local scale (5 km resolution) and accomplish this by integrating diverse satellite data sets that are related to different components of the water balance.

Integrating satellite data for estimating the different water balance components can be immensely challenging due to the varying spatial and temporal resolutions (Tamayo-Mas et al., 2018). Moreover, we lack methods to estimate several key parameters, such as recharge, GW inflow/outflow, and surface water withdrawals, to “close the loop” of the water balance and directly estimate GW withdrawals. Furthermore, existing water balance estimates have been shown, in many cases, to exhibit spatial bias, limiting their accuracy to some extent (Hashemi et al., 2017). These roadblocks have limited our ability to estimate GW withdrawals except in the special cases noted previously. In addition, traditional approaches to calibrate physical models that incorporate these diverse data sets are challenging because the models quickly become
overly complex and computationally intensive, making the parameter estimation procedure required to develop accurate estimates infeasible (Becker et al., 2019; Faunt, 2009; Moeck et al., 2018; Seibert et al., 2019; Tamayo-Mas et al., 2018).

In this research, rather than using traditional physical models, we leverage the correlations between various water balance measurements and GW withdrawals in a machine learning framework that learns the relationship between the various data sets and uses them in a predictive fashion. Here, we apply random forests (Belgiu & Drăguț, 2016; Breiman, 2001), a state of the art machine learning algorithm, to develop local scale (5 km) estimates of GW withdrawals over the entire state of Kansas. Regarding the satellite and modeled data sets, we include the water balance products from MODIS, PRISM (Parameter-elevation Regressions on Elevation Slopes Model), and land use from USDA-NASS acquired over this region between 2002 and 2019. Here, we use PRISM as it is specific to the United States, and the precipitation estimates are interpolated from a dense array of rain gauges thereby providing higher accuracies than TRMM and GPM (Cannon et al., 2017; Daly et al., 2008; Hashemi et al., 2017). We also use crop coefficients (CCs) (Allen et al., 1998) as a model predictor as an additional proxy for ET. Finally, this novel machine learning framework is validated against the annual in situ GW pumping data available over this area and the results are also compared to GRACE and GRACE-FO TWS (Landerer & Swenson, 2012; Swenson, 2012).

The remainder of this paper is divided into five main sections. We first discuss the characteristics of the study area (section 2) followed by the details of our methods (section 3). We then provide an in-depth analysis of the results (section 4) and end with a discussion of the broader implications and applicability of this work to other regions (sections 5 and 6).

2. Study Area

The High Plains Aquifer (HPA) in the central United States is one of the world's most extensive and productive aquifers helping to support agricultural demand in the United States and across the globe. It spans over eight states, and its water is predominantly used for irrigation of crops, including wheat, corn, cotton, and soybean. However, the HPA is heavily stressed (mostly in the central and southern parts) primarily due to GW overpumping driven by the large-scale demand for freshwater and increasing variations in seasonal ET and precipitation likely linked to climate change (Smidt et al., 2016). More specifically, the situation in the state of Kansas, particularly in the western area, has reached a critical point as 90% of its irrigation water is supplied by the HPA (Butler et al., 2018). Recent trends show that the water table is declining rapidly and some regions are at risk of losing their ability to pump GW. This would significantly impact the agricultural productivity of the state in the near future thereby also hindering its economic vitality (Butler et al., 2018; Deines et al., 2020).

Currently, the only feasible way to mitigate these negative impacts of extreme GW withdrawals is to reduce GW extraction (Butler et al., 2020). This is because surface water is scarce in western Kansas and the water table decline is driven by GW pumping. Over 95% of the nondomestic pumping wells in the Kansas High Plains aquifer are metered, and pumping volumes must be reported annually (Butler et al., 2018). The pumping data are available through the Water Information Management and Analysis System (WIMAS) database that can be accessed at the Kansas Geological Survey (Wilson, 2019). Since "data" are the crux of any machine learning model (Hastie et al., 2013), we chose the entire state of Kansas as our study area, which is highlighted in Figure 1. Considering the severe aquifer stress in the region and the need for proactive management solutions, we believe that this is an ideal test area for our proposed machine learning framework.

3. Methodology

The overall workflow is shown in Figure 2; the main steps are data acquisition (Step 1) using the Google Earth Engine (GEE) platform (Gorelick et al., 2017), preprocessing (Step 2), and the implementation of the machine learning model (Step 3) using Python 3. After that, we perform model evaluation and analyze the GW withdrawal predictions using the available in situ GW pumping geodatabase (GDB). Here, the input time series data consist of ET, precipitation (P), and density of the following land use types: agriculture (AGRI), urban (URBAN), and surface water (SW). We provide a detailed description of our workflow involving the different processing blocks and the time series data specifications in sections 3.1 and 3.2.
Figure 1. Map of conterminous United States highlighting the study area, Kansas, as well as the state’s five groundwater management districts (GMDs), all of which overlie the HPA.

Figure 2. Steps involved in the proposed workflow for predicting groundwater (GW) withdrawals. The machine learning model is implemented in Step 3 after data downloading and preprocessing. It is a fully automated and reproducible framework developed using open-source or freely available tools, libraries, and data. The icons were downloaded from the respective official websites and IU Digital Science Center (2013) (random forest figure). Also, QGIS (QGIS Development Team, 2019) is used for visualization.
Table 1: Specifications of the Data Sets Used in the Workflow

| Product name                                    | Time period          | Preprocessing applied                           | Brief description                                                                                                                                 |
|-------------------------------------------------|----------------------|-------------------------------------------------|---------------------------------------------------------------------------------------------------------------------------------------------------|
| MOD16A2.006: Terra net evapotranspiration 8-day  | 2002–2019 (Apr–Sep) | Temporal sum over the “ET” band                 | The ET data are available as cumulative 8-day composite images at 500 m spatial resolution (Running & Mu, 2015). The algorithm developed by Mu et al. (2011), originally based on the Penman-Monteith equation (Monteith, 1965), is used for developing these ET products. |
| global 500 m                                     |                      |                                                 |                                                                                                                                                   |
| USDA monthly spatial climate data set AN81m     | 2002–2019 (Apr–Sep) | Temporal sum over the “ppt” band                | The PRISM group provides daily and monthly gridded climate data sets for the conterminous United States wherein digital elevation models and weather station data are spatially regressed using weights derived from several physiographic factors including location and topography (Daly et al., 2008). Here, we use the monthly data sets available at 4 km spatial resolution. |
|                                                 |                      |                                                 |                                                                                                                                                   |
| USDA-NASS cropland data layers                   | 2015 (annual)        | Reclassification of the “cropland” band and     | CDL is an annual crop-specific land cover data layer (30 m spatial resolution) created for the continental United States developed using MODIS and ground-truth data (USDA-NASS, 2015). |
|                                                 |                      | Gaussian filtering                               |                                                                                                                                                   |
| Crop coefficient                                 | 2015 (annual)        | Crop coefficient values applied over CDL data   | The crop coefficient includes crop properties and soil evaporation effects and is used to estimate crop evapotranspiration. We use the midseasonal stage crop coefficient values (Allen et al., 1998) to develop a static crop coefficient raster using the USDA-NASS CDL data. |
|                                                 |                      |                                                 |                                                                                                                                                   |
| Kansas in situ groundwater pumping data          | 2002–2018 (annual)  | Groundwater pumping data to rasters             | The Kansas groundwater pumping database (GDB) is publicly available (Wilson, 2019). The GDB contains point data across all counties in Kansas.         |

3.1. Data Acquisition and Preprocessing

GEE is a cloud-based platform for large-scale geospatial data processing and analysis that is accessible to researchers and operational users (Gorelick et al., 2017). We used GEE for seamless and efficient downloading of our data sets which include MODIS (ET) and PRISM (P). The land use data (used to produce AGRI, URBAN, and SW) were downloaded directly from USDA-NASS. The specific product names along with the corresponding description are given in Table 1. Moreover, for the MOD16 and PRISM products, we restricted the time span between 1 April and 30 September (inclusive) of each year (the typical growing season; Butler et al., 2019). Our assumption was that the land cover does not change significantly between years over our study period, so we selected the USDA-NASS cropland data layer (CDL) for the year 2015 only. While many farmers rotate crops, we assume that the rotations did not significantly change the percentage of each 5 km × 5 km pixel that was cultivated over our study period. This data set does not distinguish between dryland and irrigated crops. To compensate for this, we included precipitation and actual ET as input data sets. By combining these data sets, our algorithm is able to identify pixels with low precipitation and high ET and thus high GW demand.

The Kansas in situ GW pumping data were obtained from the WIMAS database (Wilson, 2019). We discarded pixels with average extraction rates of more than 1,000 mm per year. These rates only occurred in two pixels, which had high-capacity wells that were withdrawing large volumes of water directly beneath the Missouri River and are thus more representative of surface water withdrawals than GW withdrawals. The in situ GW pumping data have not been released for 2019, and hence, this year is kept solely for forecasting.

We took the temporal sum of both the ET and P bands between April and September of each year thereby obtaining the total ET and P for that period for each year. In addition, the original USDA-NASS CDL data set (covering the conterminous United States) for the year 2015, which classifies land use based on the specific crop type, or water, urban, and so forth, was reclassified according to supporting information Table S1. Thereafter, this reclassified data set was split into separate binary rasters (AGRI, SW, and URBAN) with pixel values of 0 and 1 corresponding to the absence and presence of the respective class (OTHER class is ignored). Moreover, the original CDL data set was used to generate the CC raster by applying midseasonal stage CC values (Allen et al., 1998). The Kansas GW pumping database (GDB) was automatically converted into yearly shapefiles and subsequently rasters (5 km spatial resolution and UTM 14N projection) using...
GDAL/OGR Python APIs, GeoPandas, and Rasterio (GDAL/OGR contributors, 2019; GeoPandas developers, 2019; Gillies, 2013; Harris et al., 2020). All of these rasters (ET, P, AGRI, URBAN, SW, and CC) were reprojected to UTM 14N, resampled to 5 km spatial resolution, and clipped using the Kansas shapefile shown in Figure 1. Each 5 km × 5 km pixel contains the total annual GW withdrawal within that pixel for the years 2002–2018. The value was computed as the total volume of GW pumping in that region divided by the area, 25 km².

Furthermore, we applied a Gaussian filter available from the Scipy library (Virtanen et al., 2020) over each of the AGRI, URBAN, and SW rasters to create smoothed rasters representing the density of each land use type within a given radius, the size of which is a function of the standard deviation (σ) of the Gaussian kernel (Virtanen et al., 2020). Here, we set σ = 3 pixels for AGRI and URBAN and σ = 5 pixels for SW. The filtered data were then normalized over the interval [0, 1] where the values represent the density of these classes; that is, pixel values close to 0 have very few of the respective class within the specified window, and vice versa. Finally, the preprocessed data sets were organized into a Pandas DataFrame object (McKinney, 2010). It is noteworthy that feature ordering (order in which the predictors are provided to the model) affects model reproducibility in classification and regression trees (CART) (Breiman, 2001). This is because the splitting operation involves random feature selection to find the best split (Breiman, 1984; Pedregosa et al., 2011). In order to have consistent feature ordering, the CSV file was sorted based on the attribute names (Olson, 2015). The feature or attribute names in alphabetical order are AGRI, CC, ET, P, SW, and URBAN where the ET and P are in mm and AGRI, URBAN, and SW are measures of land type density and are unitless like CC. The GW pumping values (mm) are marked as the label set within a given radius, the size of which is a function of the standard deviation (σ) of the Gaussian kernel (Virtanen et al., 2016). Although in this work we use it for multivariate regression, the general idea remains the same where the algorithm employs a set of CART for prediction. Essentially, the random forests algorithm behaves as an ensemble or meta estimator where it uses averaging across all the CART to improve the predictive accuracy and control overfitting (Breiman, 2001). To test the performance of our model, we split the original data into training and testing data because the random forests follow a supervised machine learning approach. Here, two main hyperparameters are involved—the number of estimators or trees (n_estimators) and the number of features to consider during the best split operation (max_features). If max_features = n_features (number of actual features), then the subsample size is equal to the original input sample size where the samples are drawn with replacement. We have iterated our model using different values for both of these hyperparameters. Also, the random_state parameter controls the randomness of the sample bootstrapping and therefore remained fixed throughout the workflow for model reproducibility (Pedregosa et al., 2011).

In our final model setup, we chose n_estimators = 500 which is a commonly used value for remote sensing studies (Belgiu & Drăguț, 2016; Smith et al., 2018), random_state = 0, max_features = 6, and the other hyperparameters (maximum leaf size and minimum number of splits) as scikit-learn defaults (Pedregosa et al., 2011; scikit-learn developers, 2019). The parameters n_estimators and max_features were optimized to minimize the root-mean-square error (RMSE) on our testing (validation) data sets (described in section 4). We split the training and testing data considering a leave-one-out (year) cross-validation method where the user is able to choose a particular year as pure test and remaining years as training data respectively.

Regarding model evaluation, we considered the feature importances, partial dependence plots, and different error metrics for the predictions such as the coefficient of determination (R²), RMSE, and mean absolute error (MAE). The feature or variable importances are values (sum up to 1) signifying the impact of each variable (the higher the value, the more important the feature). Although these can be calculated in several ways, in our approach, the feature importance, also called Gini importance or mean decrease in
importance (MDI), is defined as the total decrease in node impurity (weighted by the probability of reaching a particular node) which is averaged over all the trees (Breiman, 1984). Currently, this is the most computationally efficient method and is widely used in the scientific community (Breiman, 1984; Pedregosa et al., 2011). Partial dependence plots in conjunction with the feature importances can be useful to interpret the effect of specific variables on the predicted outcome. Partial dependence plots determine the effect on the prediction outcomes when one variable is varied while accounting for all possible values from the rest (Hastie et al., 2013). The equation for partial dependence plots is as follows:

\[
\tilde{f}(x) = \frac{1}{n} \sum_{i=1}^{n} f(x, x_{ic}),
\]

where \(\tilde{f}(x)\) is the partial dependence function, \(n\) is the number of rows in the data set, \(f(x, x_{ic})\) is the random forest model, \(x\) is the predictor variable of interest, and \(x_{ic}\) denotes the values of all other variables.

This allows us to quantify how each input variable impacts GW pumping. For example, we expect regions or years with low precipitation and high ET to have high GW pumping values. If this is verified in our partial dependence plots, we can have confidence that our model is capturing the dynamics of the hydrologic system. In our proposed workflow, we use the scikit-learn library (Pedregosa et al., 2011) to obtain the different partial dependence plots which are discussed in section 4. Lastly, we also perform residual diagnostics to observe the error distribution and perform normality checks (Hastie et al., 2013).

### 4. Results and Analysis

In order to better assess our proposed model, we perform two main test cases—one with 2014 (average to slightly dry year) as a pure validation data set and the other with 2012 (extreme drought). Additionally, we also perform model evaluation based on multiple validation data sets. Here, we describe the results and the corresponding analysis for each of these scenarios.

#### 4.1. Test Case I—Year 2014

In this case, we select the preprocessed data sets for the Year 2014 as pure validation data for performing leave-one-out cross validation. The number of rows, \(n\), is 142,600 for the training data set and 8,912 for the testing or validation data set. Accordingly, we obtain an overall training score, \(R^2 \approx 0.99\). Moreover, the error metrics shown in Table 2 suggest that we are achieving good prediction results for the validation data set (or testing data set), \(R^2 \approx 0.93\), RMSE \(\approx 13.57\) mm, MAE \(\approx 4.31\) mm, and normalized MAE \(\approx 0.22\) (the MAE is normalized by the average annual pumping).

The actual and predicted GW pumping rasters for 2014 are given in Figures 3a and 3b, respectively. Next, we check the goodness of fit based on different plots (R Core Team, 2019) involving predicted values and residual diagnostics. The 1:1 relationship between the actual and predicted GW pumping values is given in Figure 3c which shows that the predictions closely follow the actual values for 2014 (\(R^2 \approx 0.93\) and RMSE \(\approx 13.50\) mm).

The model residual diagnostics are shown in Figures 4a–4d. Figure 4a shows some regions in the western portion of the study area that have pockets of high residual error. Figures 4b and 4c show the standardized residuals, obtained by dividing the residuals by their standard deviation. These should lie in the \([-2, 2]\) interval for normally distributed residuals (Hastie et al., 2013). We can see from these plots that there is a slight bias, with the residuals centered around 0 (average residual of 0.40 mm) and no significant trends in the residuals. While about 96% of the data lie within the \([-2, 2]\) interval and Figure 4b seems to follow a nearly normal distribution from the first inspection, the data do not follow a perfect normal distribution. Figure 4d shows a Q-Q plot, which can be used to determine normality. When the curve follows the horizontal line.
As observed, the curve follows a straight line between quantiles \(-2\) and \(2\), where most of the data lie, but the outliers have some skew. Some deviation from normality is expected as we are possibly overfitting the random forests model which also agrees with the lower test score \(R^2 \approx 0.93\) than training score \(R^2 \approx 0.99\). However, given that there is no significant bias or trend in the residuals, we consider the model predictions to be robust.

4.2. Test Case II—Year 2012

Similar to the 2014 data set, we selected 2012 as pure validation data set and the rest as training. Here, the model training score, \(R^2 \approx 0.99\) is also similar. The error metrics for this test case are shown in Table 3. The actual versus predicted GW pumping plots in Figures 5a and 5b for the Year 2012 indicate that even for extreme drought conditions, our approach produces satisfactory results. With \(R^2 \approx 0.84\), RMSE \(\approx 24.17\) mm, MAE \(\approx 9.72\) mm, and normalized MAE \(\approx 0.40\) this test case has significantly lower accuracy relative to Test Case I. This is caused by a lack of training data during extreme drought years, as 2012 was the driest year in our data set (Lin et al., 2017). A series of box plots showing the feature values from 2002–2018 is depicted in supporting information Figures S2a–S2d from which we can observe that 2012 is an outlier year. Thus, by holding 2012 out of the training, our random forest model had no data set from which to learn how GW pumping responds to extreme drought. We chose to hold this year out as an extreme test of our model. An \(R^2\) of 0.84 is still quite high given the extreme nature of the hold-out year, so we consider the model performance on this year as a validation of this model’s robustness.

The actual versus predicted GW pumping values for 2012 are shown in Figure 5c. As expected, there is a significantly higher scatter in Test Case II relative to Test Case I. The lower \(R^2 \approx 0.84\) and higher RMSE \(\approx 24.17\) mm for Test Case II also suggest that the predicted GW pumping values deviate more from the actual ones.
The model residual diagnostics are depicted in Figures 6a–6d. As shown in Figure 6a, there are several areas in the western portion of the study area with high residuals. Figures 6b and 6c show some bias (with an average residual value of $-2.10$ mm), and Figure 6d shows that the residuals are not normally distributed though about 96% of the standardized residuals lie in the $[-2, 2]$ interval. These observations are in concordance with the lower testing score of the random forest model, $R^2 \approx 0.84$ when compared to the high training score, $R^2 \approx 0.99$ signifying bias in the final estimates. Still, considering the extreme drought scenario, the Year 2012 is an outlier with significantly higher GW pumping than observed in the training data sets. Given these limitations, we consider the model performance for this year to be reasonable.

### 4.3. Test Case III—Using Multiple Validation Data Sets

In order to further assess our model, we held the Years 2011 to 2018 out of our model for testing ($n = 71,296$) and used the data sets from 2002–2010 for training ($n = 80,216$). Here, we only check the error metrics as given in Table 4. The overall training and testing scores for this scenario are $R^2 \approx 0.98$ and $R^2 \approx 0.80$, respectively. The test scenario had RMSE $\approx 22.52$ mm, MAE $\approx 8.18$ mm, and normalized MAE $\approx 0.41$. The residual diagnostics for Test Case III are provided in supporting information Figure S1 which shows that the residuals for the mean error raster are almost normally distributed with a mean and standard deviation of $-0.07$ and $17.17$ mm, respectively. The $R^2$ in the training data set is significantly higher than the $R^2$ of the testing data set. This can be a sign of model overfit, when models fit noise rather than causal mechanisms to improve the match of training data, at the expense of true model predictive accuracy (Lever et al., 2016). However, in this case the disparity between training and testing scores is caused primarily by the difference in climate during the training and testing periods. The testing period...
was intentionally selected to include two major drought years (2011 and 2012), which both had ET and P values outside the range of the training data (supporting information Figures S2a and S2b). In spite of the discrepancy between training and testing scores, the $R^2$ value of 0.80 indicates a reasonably good fit even in years where the testing data falls outside of the range the training data experienced, indicating a model that is robust in predicting pumping during extreme events.

5. Discussion

We chose the random forest model developed during Test Case I as the final one for the GW pumping prediction, as it used the most training data and was able to learn from both wet and dry years and is most likely to be effective for future predictions. After the model evaluation phase, we observed that the feature importances in all three test cases were similar, with the most important features being the land use classes, that is, AGRI, SW, and URBAN. This finding is in agreement with GW withdrawals predominantly occurring in agricultural areas with limited surface water availability (Butler et al., 2018). More specifically, the feature importances (rounded to two decimal places) of AGRI, SW, URBAN, ET, P, and CC are 0.33, 0.27, 0.19, 0.09, 0.08, and 0.03 in descending order, respectively. This metric gives higher importance to spatially variable but temporally static predictors (i.e., land use) than spatiotemporal variable predictors (i.e., P and ET). The main reason for this is that GW withdrawals vary more spatially than they do temporally, so spatial predictors explain most of the variance in GW pumping values. However, this metric does not provide a complete picture of variable importance, as P and ET are critical

| Year | $R^2$ | RMSE (mm) | MAE (mm) | Normalized MAE |
|------|-------|-----------|----------|---------------|
| 2002 | 0.98  | 9.28      | 3.67     | 0.14          |
| 2003 | 0.99  | 6.56      | 2.36     | 0.10          |
| 2004 | 0.99  | 5.65      | 1.88     | 0.10          |
| 2005 | 0.99  | 4.48      | 1.62     | 0.09          |
| 2006 | 0.99  | 5.89      | 2.14     | 0.10          |
| 2007 | 0.98  | 6.63      | 2.59     | 0.14          |
| 2008 | 0.99  | 5.34      | 1.75     | 0.09          |
| 2009 | 0.99  | 4.49      | 1.49     | 0.08          |
| 2010 | 0.99  | 4.86      | 1.76     | 0.09          |
| 2011 | 0.98  | 9.49      | 3.34     | 0.13          |
| 2012 | 0.84  | 24.17     | 9.72     | 0.40          |
| 2013 | 0.99  | 6.18      | 1.95     | 0.09          |
| 2014 | 0.99  | 5.04      | 1.60     | 0.08          |
| 2015 | 0.99  | 4.35      | 1.48     | 0.08          |
| 2016 | 0.99  | 4.61      | 1.65     | 0.10          |
| 2017 | 0.98  | 5.63      | 1.70     | 0.10          |
| 2018 | 0.99  | 4.67      | 1.68     | 0.10          |

Note. The bold ones show the result considering only the test data for three test cases.
variables for quantifying any temporal variations in GW, including the effect of drought or wet years, which is one of the key goals of this study. To improve our understanding of the importance of spatial and spatio-temporal variables, we compared the temporal variations in GW withdrawals from 2002–2018 considering first spatial predictors with no temporal variance (AGRI, SW, URBAN, and CC) and then predictors with spatiotemporal variance (ET and P) separately. Supporting information Figures S3a–S3e indicate that using only ET and P (feature importances of 0.49 and 0.51, respectively) results in an overall underprediction of the mean annual GW withdrawals in each groundwater management district (GMD). However, there is a substantial overprediction for the year 2012 when the entire state of Kansas (supporting information Figure S3f) is considered. This could be attributed to the high GW withdrawals in 2012 (drought year) which lead to significant model overfitting in the absence of land use predictors. Furthermore, when only the spatial predictors are used for model training, we do not observe any temporal variations which are expected. In this case, the feature importances of AGRI, SW, URBAN, and CC are 0.40, 0.31, 0.25, and 0.04 in descending order, respectively. Therefore, the best predictions are obtained when both the spatial and spatiotemporal predictors are synergistically introduced during the model training.

To further explore the importance of each variable, we employ partial dependence plots which indicate a plausible agreement between the model behavior and the expected heuristic relationships (Figures 7a–7e). Groundwater pumping increases with higher AGRI and ET values and decreases with increased P and SW, as would be expected. Also, the GW pumping varies significantly with the URBAN predictor. This could be
attributed to the variability in the GW demand of urban areas in Kansas. CC (which has the lowest feature importance) shows a prominent increase in partial dependence (GW) between 1.15 and 1.20 which is the midseasonal CC range of wheat, corn, soy, sorghum, and cotton—the most common agricultural crops in Kansas. It is noteworthy that the land use predictors and the CC are not heavily correlated with each other. However, SW has moderate correlation with ET and P, and ET and P are strongly correlated. Although the partial dependence function assumes that the features are uncorrelated (Hastie et al., 2013), Altmann et al. (2020) show that random forest partial dependence is somewhat robust toward correlated features. The correlation matrix diagram shown in supporting information Figure S2e can be used for obtaining further details regarding feature correlation.

While our model was reasonably accurate, it did display bias in some regions (Figures 4a, 6a, and supporting information Figures S1a–S1i), particularly in west central Kansas, where we are overpredicting pumping, most notably during the major drought year of 2012 (Figure 6a). This could be related to the transmissivity of the aquifer—the aquifer thickness in many of the areas with overpredicted pumping has reached a point that it can no longer support large-scale pumping for irrigated agriculture (Butler et al., 2018). Thus, these areas have high GW demand due to the drought but do not have sufficient resources to meet the demand. We would expect those areas to have lower ET, but the dependence of pumping on ET resulting from the random forests may not be great enough to identify those areas.

The model we developed provides reasonable estimates of GW pumping in Kansas. Since the model was trained on climate (i.e., precipitation and ET) data and land use patterns specific to that region, the relationships it learned between those predictor variables and GW pumping estimates will likely hold in other regions with similar climates and land use patterns. However, it is less transferrable to regions with significantly different climates or land use patterns.

5.1. Temporal Trends in Groundwater Pumping and TWS

As a final step of our model evaluation, we used Test Case III (2011–2018 as pure test data) to compare (Figure 8) the GW pumping values (both actual and predicted) for the entire study area and three specific GMDs (GMD 3, 4, and 5) to the GRACE and GRACE-FO TWS measurements (Landerer & Swenson, 2012; Swenson, 2012). GMDs 3, 4, and 5 cover southwest, northwest, and south central regions of Kansas (Butler et al., 2016) and are under significant stress due to heavy GW pumping. GMDs 3 and 4 are in a semi-arid climate, while GMD5 is in a subhumid climate. Due to their varying climate and high water demand,
Figure 8. Actual versus predicted mean annual groundwater (GW) pumping for (a) GMD 3, (b) GMD 4, (c) GMD 5, and (d) entire Kansas. Monthly GRACE/GRACE-FO TWS between 2002 and 2019 (e) are included for comparison with the entire Kansas plot. Here, Years 2011–2018 are test data (Test Case III) and forecasts are made for 2019.
these GMDs have conditions similar to a range of other aquifers that are under heavy water stress, so the performance of our model over these regions is indicative of model performance over regions with similar conditions globally.

From Figure 8a, we observe that the GW pumping predictions for GMD 3 are lower than the actual values. However, the predictions for GMD 4 and 5 match quite well with the actual pumping measurements (Figures 8b and 8c). GMD 3 is likely underpredicted because, in general, the actual values of pumping in GMD 3 are the highest of the study area. Random forests tend to produce conservative estimates that lean toward the mean and thus underpredict and overpredict the highest and lowest estimates, respectively. However, the predictions are still reasonably close, and the other GMD in a semiarid climate (GMD 4) matches closely with observed withdrawals. Thus, we consider the model to perform well in both semiarid and subhumid climates.

Since the TWS measurements in Figure 8e are too coarse to compare at the GMD scale, we use the GW pumping estimates from the entire state of Kansas (Figure 8d) for the comparison. The pattern for the mean annual GW pumping values for Kansas is in general agreement with that for the TWS values. Note that GW pumping most closely matches the change in TWS during each growing season (spring to fall) for any given year. Essentially, this highlights the effectiveness and robustness of our model even when we leave out several years of data from training.

Many studies have found data from the GRACE to be a strong predictor of changes in GW storage at regional scales, that is, hundreds of km (Landerer & Swenson, 2012; Rodell, 2004; Rodell et al., 2007, 2009; Tiwari et al., 2009). For this reason, we initially included GRACE TWS as a predictor in our model. However, we found that our model performed better without GRACE data, so it was ultimately removed. We believe that the value of GRACE in GW storage changes is diminished in our study both due to the coarse GRACE resolution (400 km) (Miro & Famiglietti, 2018) relative to the resolution of our model (5 km), and also due to its correlation with precipitation, which is another strong indicator of drought in Kansas. We consider it likely that GRACE would be a more useful predictor for extending assessments to broader regions, where GRACE is a more consistent indicator of drought than precipitation, which has different normal levels for different regions.

5.2. Method Transferability to Similar Regions

In order to verify the transferability of our approach to other regions having similar spatiotemporal characteristics, we built a new test case where we explicitly held out a large region as test data and retrained the model using the remaining areas. Essentially, this is a different train-test split approach (compared to Test Cases I–III) where we consider all the annual data (2002–2018) for model training except those belonging to the test region. More specifically, we selected the test area shown in supporting information Figure S4a because the water availability varies from east to west (less water going west) and north to south (less water going south). This region also had land use and climate patterns that were within the bounds of the other training data covering all the years (2002–2018). The mean annual GW withdrawal predictions corresponding to the test area (supporting information Figures S4b–S4e) indicate that our method was able to predict GW withdrawals reasonably well, having $R^2 \approx 0.23$, $\text{MAE} \approx 16.01$ mm, $\text{RMSE} \approx 31.51$ mm, and normalized $\text{MAE} \approx 0.84$. Although the $R^2$ is low relative to Test Cases I–III, the broader spatial trends of GW pumping, as well as the temporal trends, are captured. The lower fit for this data set is expected, since the entire region being tested has not been “seen” by our model. Thus, while our model is most accurate in predicting GW pumping within the study area and could be used for predicting future GW pumping or filling in temporal gaps, it may also be used to make first-order estimates of GW pumping in other regions with similar climate and land use patterns.

5.3. Uncertainty Associated With MOD16, PRISM, and CDL

We selected MOD16 because it is the highest resolution (500 m) globally available ET product. Moreover, various studies have shown that the ET estimates from MOD16 perform reasonably well when validated against eddy covariance flux tower data (Cleugh et al., 2007; Lakshmi et al., 2018; Running et al., 2019; Tang et al., 2011; Velpuri et al., 2013). More specifically, the validation test against individual AmeriFlux towers shows about 10–30% uncertainty associated with the MOD16 ET products (Mu et al., 2011; Velpuri et al., 2013). Furthermore, all remote sensing-based ET products will have certain assumptions leading to
Acknowledgments
We would like to acknowledge all the open-source software and data communities for making their resources publicly available. Particularly, we would like to thank NASA/JPL (https://grace.jpl.nasa.gov/data-analysis-tool/ and https://doi.org/10.5067/MODIS/MOD16A2.006), USDA (https://nassgeodata.gmu.edu/CropScape/), and the PRISM group (http://www.prism.oregonstate.edu/) for providing all the necessary remote sensing data sets. The work of the third author was supported, in part, by the Kansas Water Plan under the Ogallala-High Plains Aquifer Assessment Program (OHPAAP) and the United States Department of Agriculture (USDA) and the United States National Science Foundation (NSF) under USDA-NIFA/NSF INFEWS subaward RC108063UK. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the OHPAAP, USDA, or NSF. We thank Brownie Wilson of the Kansas Geological Survey for helping us access the WIMAS database (http://hercules.kgs.ku.edu/geohydro/wimas/). Finally, we are grateful to our colleagues and families for their continuous motivation and support.

6. Conclusions
In this research, we developed a new machine learning framework for predicting GW withdrawals at a local scale (5 km spatial resolution). This holistic approach integrates various openly available satellite products that are related in different ways to GW withdrawals. These include ET (MODIS), precipitation (PRISM), and land use data (USDA-NASS). We deployed a machine learning framework using random forests to build a self-learnable method that incorporates these data sets along with in situ GW extraction data (Wilson, 2019). Our workflow is fully automated and has been implemented using open-sourced or freely available software tools and libraries.

A thorough model assessment showed that our proposed approach works satisfactorily with predictions having high $R^2$ values and low RMSE and MAE. In this work, we considered three test scenarios to validate our method, two of which included the extreme drought period of 2012. Moreover, the random forest feature importance agrees with the expected dependencies of GW withdrawals.

As climate change, dietary shifts, and population growth increase the global stress on GW resources, it is critical to develop sustainable GW management practices, yet the volume of GW withdrawals, a key quantity for such efforts, is poorly constrained in many regions. Although one of us has repeatedly called for greater monitoring of GW withdrawals (Butler et al., 2016, 2018, 2020), that has proven challenging in many areas. This method enables water managers to predict GW withdrawals from anthropogenic and climate drivers—such knowledge could lead to proactive implementation of sustainable solutions related to GW withdrawal practices.

Conflict of Interest
We confirm that there is no conflict of interest among the authors of this manuscript.

Data Availability Statement
The entire project code along with the preprocessed data sets is available at GitHub (https://github.com/compstat-iml/HydroMST).
Becker, R., Koppa, A., Schulz, S., Usman, M., aus der Beek, T., & Schüth, C. (2019). Spatially distributed model calibration of a highly managed hydrological system using remote sensing-derived ET data. *Journal of Hydrology*, 577, 123944. https://doi.org/10.1016/j.jhydrol.2019.123944

Belgu, M., & Drąguț, L. (2016). Random forest in remote sensing: A review of applications and future directions. *ISPRS Journal of Photogrammetry and Remote Sensing*, 114, 24–31. https://doi.org/10.1016/j.isPRS.2016.01.011

Boryan, C., Yang, Z., Mueller, R., & Craig, M. (2011). Monitoring US agriculture: The US Department of Agriculture, National Agricultural Statistics Service, Cropoland Data Layer Program. *Geocarto International*, 26(5), 341–358. https://doi.org/10.1080/1010049.2011.562309

Breiman, L. (2001). Random forests. *Machine Learning*, 45(1), 5–32. https://doi.org/10.1023/A:1010933404324

Breiman, L. (1984). *Classification and regression trees* (1st ed.). New York, USA: Routledge. https://doi.org/10.1007/978-1315319470

Butler, J. J., Bohling, G. C., Whittemore, D. O., & Wilson, B. B. (2020). A roadblock on the path to aquifer sustainability: Underestimating the impact of pumping reductions. *Environmental Research Letters*, 15(1), 014003. https://doi.org/10.1088/1748-9326/ab6002

Butler, J. J., Whittemore, D. O., Reboulet, E., Knobbe, S., Wilson, B. B., & Bohling, G. C. (2019). *High plains aquifer index well program: 2018 Annual Report*. Lawrence, Kansas, USA: Kansas Geological Survey. http://www.ksge.ksu.edu/hsy/annualPub/2019/OPRF19_19_OFR2019-19.pdf

Butler, J. J., Whittemore, D. O., Wilson, B. B., & Bohling, G. C. (2016). A new approach for assessing the future of aquifers supporting irrigated agriculture. *Geophysical Research Letters*, 43, 2004–2010. https://doi.org/10.1002/2016GL067879

Butler, J. J., 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 Resources Research*, 46, 815–828. https://doi.org/10.1002/2018WR022060.2018.151556

Cannon, F., Ralph, F. M., Wilson, A. M., & Lettenmaier, D. P. (2017). GPM satellite radar measurements of precipitation and freezing level in atmospheric rivers: Comparison with ground-based radars and Reanalyses. *Journal of Geophysical Research: Atmospheres*, 122, 12,747–12,764. https://doi.org/10.1002/2017JD027355

Chan, S. K., Bindlish, R., O’Neill, P. E., Njoku, E., Jackson, T., Collinder, A., et al. (2016). Assessment of the SMAP passive soil moisture product. *IEEE Transactions on Geoscience and Remote Sensing*, 54(8), 4994–5007. https://doi.org/10.1109/TGRS.2016.2561938

Chauward, E., & Farr, T. G. (2019). A new method for isolating elastic from inelastic deformation in aquifer systems: Application to the San Joaquin Valley, CA. *Geophysical Research Letters*, 46, 10,800–10,809. https://doi.org/10.1029/2019GL084418

Cleugh, H. A., Leuning, R., Mu, Q., & Running, S. W. (2007). Regional evaporation estimates from meteorological temperature and precipitation across the conterminous United States. *Remote Sensing of Environment*, 106(3), 285–304. https://doi.org/10.1016/j.rse.2006.07.007

Daly, C., Halbleib, M., Smith, J. I., Gibson, W. P., Doggett, M. K., Taylor, G. H., et al. (2008). Physiographically sensitive mapping of climatological temperature and precipitation across the conterminous United States. *International Journal of Climatology*, 28(15), 2031–2064. https://doi.org/10.1002/joc.1688

Deines, J. M., Kendall, A. D., & Hyndman, D. W. (2017). Annual irrigation dynamics in the U.S. Northern High Plains derived from Landsat satellite data. *Geophysical Research Letters*, 44, 9350–9360. https://doi.org/10.1002/2017GL073471

Deines, J. M., Schipanski, M. E., Golden, B., Zipper, S. C., Nozari, S., Rottler, C., et al. (2020). Transitions from irrigated to dryland agriculture in the Ogallala aquifer: Land use suitability and regional economic impacts. *Agricultural Water Management*, 233, 106,061. https://doi.org/10.1016/j.agwat.2020.106061

Dogrul, E. C., Schmid, W., Hanson, R. T., Kadir, T., & Chung, F. (2016). Integrated water flow model and modflow-farm process: A comparison of theory, approaches, and features of two integrated hydrologic models. In *California Natural Resources Agency*. Sacramento, CA: California Department of Water Resources. http://pubs.er.usgs.gov/publication/7017848

Erban, L. E., Gorlick, S. M., Zekker, H. A., & Fendorf, S. (2013). Release of arsenic to deep groundwater in the Mekong Delta, Vietnam, linked to pumping-induced land subsidence. *Proceedings of the National Academy of Sciences*, 110(34), 13,753–13,756. https://doi.org/10.1073/pnas.1300501110

Famiglietti, J. S., Lo, M., Ho, S. L., Bethune, J., Anderson, K. J., Syed, T. H., et al. (2011). Satellites measure recent rates of groundwater depletion in California’s Central Valley. *Geophysical Research Letters*, 38, L03403. https://doi.org/10.1029/2010GL046442

Faunt, C. C. (Ed.) (2009). *Groundwater availability of the Central Valley Aquifer, California*. In U.S. Geological Survey Professional Paper (Vol. 1766, pp. 59–116). Sacramento, CA: U.S. Geological Survey. https://doi.org/10.3133/pp1766

Frappart, F., & Bourrel, L. (Eds) (2018). *The use of remote sensing in hydrology*. Basel, Switzerland: MDPI. https://doi.org/10.3390/books978-3-03482-910-4

Galloway, D. L., & Burbye, T. J. (2011). Review: Regional land subsidence accompanying groundwater extraction. *Hydrogeology Journal*, 19(8), 1459–1486. https://doi.org/10.1007/s10040-011-0775-5

GDAL/OGR contributors. (2019). *GDAL/OGR GeoSpatial Data Abstraction software library*. Open Source Geospatial Foundation. https://gdal.org

Geopandas developers. (2019). *GeoPandas 0.6.0*. http://geopandas.org/

Gillies, S. (2013). *Rasterio: Geospatial raster I/O for Python programmers*. Mapbox. https://github.com/mapbox/rasterio

Gleeson, T., Villholth, K., Taylor, R., Perrone, D., & Hyndman, D. (2019). Groundwater: A call to action. *Nature*, 578(7786), 213–213. https://doi.org/10.1038/d41586-019-03711-0

Gorick, N., Hancher, M., Dixon, M., Ilyushchenko, S., Thau, D., & Moore, R. (2017). Google Earth Engine: Planetary-scale geospatial analysis for everyone. *Remote Sensing of Environment*, 202, 18–27. https://doi.org/10.1016/j.rse.2017.06.031

Harris, C. R., Millman, K. J., van der Walt, S. J., Gommers, R., Virtanen, P., Cournapeau, D., et al. (2020). Array programming with NumPy. *Trelgol Publishing*. Retrieved from https://books.google.com/books?id=tKuSGzAACAAJ

Hashemi, H., Nordlin, M., Lakshmi, V., Huffman, G. J., & Knight, R. (2017). Bias correction of long-term satellite monthly precipitation product (TRMM 3B43) over the conterminous United States. *Journal of Hydrometeorology*, 18(9), 2491–2509. https://doi.org/10.1175/JHM-D-17-0025.1

Hastie, T., Tibshirani, R., & Friedman, J. (2013). *The elements of statistical learning: Data mining, inference, and prediction*. New York, USA: Springer. https://books.google.com/books?id=yPIZBwAAQBAJ

Hoffmann, J., Zeckeb, H. A., Galloway, D. L., & Ameling, F. (2001). Seasonal subsidence and rebound in Las Vegas Valley, Nevada, observed by synthetic aperture radar interferometry. *Water Resources Research*, 37(6), 1551–1566. https://doi.org/10.1029/2000WR900404

IU Digital Science Center. (2013). *Harp random forests*. Bloomington, Indiana: IU Digital Science Center. https://disc-spidal.github.io/harp/docs/examples/el/

Lakshmi, V., Fayne, J., & Bolten, J. (2018). A comparative study of available water in the major river basins of the world. *Journal of Hydrology*, 567, 510–532. https://doi.org/10.1016/j.jhydrol.2018.10.038
Landreer, F. W., & Swenson, S. C. (2012). Accuracy of scaled GRACE terrestrial water storage estimates. Water Resources Research, 48, W04531. https://doi.org/10.1029/2011WR011453

Leidner, A. K., & Buchanan, G. M. (Eds.). (2018). Satellite remote sensing for conservation action. Edinburgh, UK: Cambridge University Press. https://doi.org/10.1017/9781108631129

Lever, K., Krzywinski, M., & Altman, N. (2016). Model selection and overfitting. Nature Methods, 13(9), 703–704. https://doi.org/10.1038/nmeth.3968

Lin, X., Harrington, J., Ciampietti, I., Gowda, P., Brown, D., & Kisekka, I. (2017). Kansas trends and changes in temperature, precipitation, drought, and frost-free days from the 1890s to 2015. Journal of Contemporary Water Research & Education, 162(1), 16–30. https://doi.org/10.1111/j.1936-704X.2017.03257.x

MardanDoost, B., Brook, A. E., Bledova, O., & Barik, R. (2013). Estimating irrigation demand with geospatial and in-situ data: Application to the high plains aquifer, Kansas. USA. Agricultural Water Management, 123, 105,675. https://doi.org/10.1016/j.agwat.2019.06.010

Margat, J., & van der Gun, J. (2013). Groundwater around the world: A geographic synopsis (1st ed.). London, UK: CRC Press. https://www.routledge.com/Groundwater-around-the-World-A-Geographic-Synopsis/Margat-Gun/p/book/9781138000346

Mckinney, W. (2010). Data structures for statistical computing in Python. In S. van der Walt & J. Millman (Eds.), Proceedings of the 9th Python in Science Conference (SciPy 2010) (pp. 51–56). Austin, TX: SciPy Conference. https://conference.scipy.org/proceedings/scipy2010/pdfs/mckinney.pdf

McShane, R. R., Driocoll, K. P., & Sando, R. (2017). A review of surface energy balance models for estimating actual evapotranspiration with remote sensing at high spatiotemporal resolution over large extents (Scientific Investigations Report 2017–5087). https://www.fs.fed.us/rm/pubs_journals/2017/irms_2017/irms_2017_mchshe_001.pdf

Miro, M., & Famiglietti, J. (2018). Downscaling GRACE remote sensing datasets to high-resolution groundwater storage change maps of California’s Central Valley. Remote Sensing, 10(1), 143. https://doi.org/10.3390/rs10010143

Moeck, C., von Freyberg, J., & Schirmer, M. (2018). Groundwater recharge predictions in contrasted climate: The effect of model complexity and calibration period on recharge rates. Environmental Modelling & Software, 103, 74–89. https://doi.org/10.1016/j.envsoft.2018.02.005

Monteith, J. L. (1965). Evaporation and the environment. Symposium of the Society for Experimental Biology, 19, 205–234. https://repository.rothamsted.ac.uk/download/8ac29c1e04f617575d8be9642ec6556306f3a39bb584a1eb212ff44a1e13879831/Monteith65.pdf

Mouritzinis, S., Rattalino Edreira, J. I., Conley, S. P., & Grassini, P. (2017). From grid to application example in the continental US. Remote Sensing of Environment, 118(8), 163–172. https://doi.org/10.1016/j.rse.2016.10.013

Mu, Q., Zhao, M., & Running, S. W. (2011). Improvements to a MODIS global terrestrial evapotranspiration algorithm. Remote Sensing of Environment, 115(8), 1781–1800. https://doi.org/10.1016/j.rse.2011.02.019

Nie, W., Dietrich, B. R., Rodell, M., Kumar, S. V., Anderson, M. C., & Hain, C. (2018). Groundwater withdrawals under drought: Reconciling GRACE and land surface models in the United States High Plains aquifer. Water Resources Research, 54, 5282–5299. https://doi.org/10.1029/2017WR022179

Olson, R. (2015). Feature order affects the performance of tree-based classifiers. Github. https://github.com/scikit-learn/scikit-learn/issues/5394

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

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., et al. (2011). Scikit-learn: Machine learning in Python. Journal of Machine Learning Research, 12, 2825–2830. http://www.jmlr.org/papers/volume12/pedregosa11a/pedregosa11a.pdf

QGIS Development Team (2019). QGIS Geographic Information System. Open Source Geospatial Foundation Project. http://qgis.osgeo.org/

R Core Team (2019). R: A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing. https://www.r-project.org/

Roddell, M. (2004). Basin scale estimates of evapotranspiration using GRACE and other observations. Geophysical Research Letters, 31, L20304. https://doi.org/10.1029/2004GL020873

Roddell, M., Chen, J., Kato, H., Famiglietti, J. S., Nigro, J., & Wilson, C. R. (2007). Estimating groundwater storage changes in the Mississippi River basin (USA) using GRACE. Hydrogeology Journal, 15(1), 159–166. https://doi.org/10.1007/s10040-006-0103-7

Roddell, M., Velicogna, I., & Famiglietti, J. S. (2009). Satellite-based estimates of groundwater depletion in India. Nature, 460(7258), 999–1002. https://doi.org/10.1038/nature08238

Running, S. W., & Mu, Q. (2015). MOD16A2 MODIS/Terra Evapotranspiration 8-day L4 Global 500m SIN Grid. Missoula, Montana, USA: NASA LP DAAC. https://doi.org/10.5067/MODIS/MOD16A2.006

Roznikov, S. W., Mu, Q., Zhao, M., & Moreno, A. (2019). MODIS global terrestrial evapotranspiration (ET) product (MOD16A2/A3 and year-end gap-filled MOD16A2GF/A3GF) NASA Earth Observing System MODIS Land Algorithm (for Collection 6). Missoula, Montana, USA: NASA LP DAAC. https://modis-land.gac.nasa.gov/pdf/MOD16UsersGuideV2.02019.pdf

Schmid, W. (2004). A farm package for MODFLOW-2000: Simulation of irrigation demand and conjunctively managed surface-water and ground-water supply (PhD Dissertation), Tucson, AZ. Department of Hydrology and Water Resources, The University of Arizona. https://repository.arizona.edu/handle/10150/191271

Schneider, S. H., Root, T. L., & Mastrandrea, M. D. (Eds.). (2011). Wrangling climate and weather (Vol. 1, 2nd ed., pp. 250–270). New York: Oxford University Press. https://doi.org/10.1093/acred/9780199763524.001.0001

scikit-learn developers. (2019). 3.2.4.3.2. sklearn.ensemble. RandomForestRegressor — scikit-learn 0.23.1 Documentation. https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestRegressor.html

Selbert, J., Staudinger, M., & van Meerfeld, H. J. (2019). Validation and over-parameterization—Experiences from hydrological modeling. In C. Beisbart & N. Saam (Eds.), Computer Simulation Validation: Simulation Foundations, Methods and Applications (pp. 811–834). Cham, Switzerland: Springer. https://doi.org/10.1007/978-3-319-70766-2_33

Smajgl, A., Ward, J., & Pluschke, L. (2016). The water-energy Nexus—Realising a new paradigm. Journal of Hydrology, 533, 533–540. https://doi.org/10.1016/j.jhydrol.2015.12.033

Smith, S. J., Haack, E. M. K., Kendall, A. D., Deines, J. M., Pei, L., Cotterman, K. A., et al. (2016). Complex water management in modern agriculture: Trends in the water-energy-food nexus over the High Plains Aquifer. Science of the Total Environment, 566–567, 988–1001. https://doi.org/10.1016/j.scitotenv.2016.05.127

Smith, R., Knight, R., Chen, J., Reeves, J. A., Zebker, H. A., Farr, T., & Liu, Z. (2017). Estimating the permanent loss of groundwater storage in the southern San Joaquin Valley, California. Water Resources Research, 53, 2133–2148. https://doi.org/10.1002/2016WR021961
Smith, R., Knight, R., & Fendorf, S. (2018). Overpumping leads to California groundwater arsenic threat. *Nature Communications*, 9(1), 2089. https://doi.org/10.1038/s41467-018-04475-3

Smith, R. G., & Majumdar, S. (2020). Groundwater storage loss associated with land subsidence in Western United States mapped using machine learning. *Water Resources Research*, 56, e2019WR026621. https://doi.org/10.1029/2019WR026621

Swenson, S. C. (2012). *GRACE monthly land water mass grids NetCDF Release 5.0*. Ver. 5.0. CA, USA: PO.DAAC. https://doi.org/10.5067/TELND-NC005

Tamayo-Mas, E., Bianchi, M., & Mansour, M. (2018). Impact of model complexity and multi-scale data integration on the estimation of hydrogeological parameters in a dual-porosity aquifer. *Hydrogeology Journal*, 26(6), 1917–1933. https://doi.org/10.1007/s10040-018-1745-y

Tang, R., Li, Z.-L., & Chen, K.-S. (2011). Validating MODIS-derived land surface evapotranspiration with in situ measurements at two AmeriFlux sites in a semiarid region. *Journal of Geophysical Research*, 116, D04106. https://doi.org/10.1029/2010JD014543

Tiwari, V. M., Wahr, J., & Swenson, S. (2009). Dwindling groundwater resources in northern India, from satellite gravity observations. *Geophysical Research Letters*, 36, L18401. https://doi.org/10.1029/2009GL039401

USDA-NASS (2015). *Published crop-specific data layer*. Washington, DC: USDA NASS. https://nassgeodata.gmu.edu/CropScape/

USDA-NASS (2016). *Kansas Cropland Data Layer*NASS/USDA (p. 2015). Washington, DC: USDA, NASS. https://www.nass.usda.gov/Research_and_Science/Cropland/metadata/metadata_ks15.htm

Velpuri, N. M., Senay, G. B., Singh, R. K., Bohms, S., & Verdin, J. P. (2013). A comprehensive evaluation of two MODIS evapotranspiration products over the conterminous United States: Using point and gridded FLUXNET and water balance ET. *Remote Sensing of Environment*, 139, 35–49. https://doi.org/10.1016/j.rse.2013.07.013

Virtanen, P., Gommers, R., Oliphant, T. E., Haberland, M., Reddy, T., Cournapeau, D., et al. (2020). SciPy 1.0: fundamental algorithms for scientific computing in Python. Retrieved from http://www.scipy.org/

Wilson, B. B. (2019). *Water Information Management and Analysis System (WIMAS) for the Web*. Lawrence, Kansas, USA: Kansas Geological Survey. http://hercules.kgs.ku.edu/geohydro/wimas/query_setup.cfm

Yi, Z., Zhao, H., & Jiang, Y. (2018). Continuous daily evapotranspiration estimation at the field-scale over heterogeneous agricultural areas by fusing ASTER and MODIS data. *Remote Sensing*, 10(11), 1694. https://doi.org/10.3390/rs10111694