Leveraging statistical learning theory to characterize the U.S. water consumption

Ellen Wongso[1], Roshanak Nateghi[1], Benjamin Zaitchik[2], Steven Quiring[3], and Rohini Kumar[4]

[1]School of Industrial Engineering, Purdue University, West Lafayette, IN, USA
[2]Department of Earth and Planatery Sciences, Johns Hopkins University, Baltimore, MD, USA
[3]Department of Geography, Ohio State University Columbus OH, USA
[4]UFZ-Helmholtz Centre for Environmental Research, Leipzig, Germany

November 24, 2022

Abstract

Access to accurate estimates of water withdrawal is requisite for urban planners as well as operators of critical infrastructure systems to make optimal operational decisions and investment plans to ensure reliable and affordable provisioning of water. Furthermore, identifying the key predictors of water withdrawal is important to regulators for promoting sustainable development policies to reduce water use. In this paper, we developed a rigorously evaluated predictive model, using statistical learning theory, to estimate state-level, per-capita water withdrawal as a function of various geographic, climatic and socio-economic variables. We then harnessed the data-driven predictive model to identify the key factors associated with high water-usage intensity among different sectors in the U.S. We analyzed the predictive accuracy of a range of parametric models (e.g., generalized linear models) and non-parametric, flexible learning algorithms (e.g., generalized additive models, multivariate adaptive regression splines and random forest). Our results identified irrigated farming, thermo-electric energy generation and urbanization as the most water-intensive anthropogenic activities, on a per-capita basis. Among the climate factors, precipitation was also found to be a key predictor of per-capita water withdrawal, with drier conditions associated with higher water withdrawals. Results of the first-order sensitivity analysis indicated changes between +/-10% in the future water withdrawal across the U.S., in response to precipitation changes, by the end of the 21st Century under the business-as-usual scenario. Overall, our study highlights the utility of leveraging statistical learning theory in developing data-driven models that can yield valuable insights related to the water withdrawal patterns across expansive geographical areas.

Leveraging statistical learning theory to characteri e the U.S. water consumption

[1] [1] [2] [3] [4]

- U.S.
-
- U.S. (±10%)
  RCP8.5

Corresponding author: Rohini Kumar; Roshanak Nateghi, rohini.kumar@ufz.de;rnateghi@purdue.edu

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

$$= \left[- \hat{ }( ) \right]^2$$
$$= [- ( )]^2 + \left[\hat{ }( ) - ( ) \right]^2 + \left[\hat{ }( ) - \hat{ }( ) \right]$$
$$= ( ) + {}^2 + \hat{ }( ) \qquad (1)$$

Th[illegible]

(e.g., ... U.S. ... (e.g., ... (e.g., ...

$= ( ) + ;$ ... $( )$ ... (i.e., ...

$$= \int ( ) \hat{ }( ) ( ) \qquad (2)$$

$( )$ ...

(Hastie, Tibshirani, & Friedman, 2009).

(e.g., ... (e.g., ... (GAM), ... (MARS) ... (RF)) ... (RF) ...

(2001), ...

$$(\ ) = \frac{1}{\ } \sum_{i=1}^{m_{tree}} {}_{i}(\ ) \qquad (3)$$

[illegible] $_i$ [illegible] — [illegible] (e.g., [illegible]). T[illegible] [illegible] ([illegible]) [illegible] $_i$ [illegible] T[illegible] ([illegible]) [illegible] ( ) [illegible]

R[illegible] (H[illegible], 2009). T[illegible] T[illegible] T[illegible] T[illegible] T[illegible] (OOB) [illegible] (i.e., [illegible]) [illegible]

$$= \frac{1}{\ } \sum_{i=1}^{n} (_{i} - \hat{}_{i}')^2 \qquad (4)$$

[illegible] $\hat{}_i'$ [illegible] OOB [illegible] th [illegible] (L[illegible] & W[illegible], 2002). S[illegible] (PDPs) [illegible] PDP [illegible] $_j$ [illegible] (i.e., [illegible]). M[illegible] PDP [illegible] (H[illegible] [illegible], 2009):

$$(\hat{}_j)(_j) = \frac{1}{\ } \sum_{i=1}^{n} (\hat{}_j)(_j \quad _{-j;i}) \qquad (5)$$

[illegible] $\hat{}_j$ [illegible] ; [illegible] (i.e., [illegible]) ; [illegible] $_{-j}$ [illegible] $_j$. T[illegible] PDP [illegible] $_{-j}$ [illegible] $\hat{}$ [illegible] $_j$ is [illegible] $_{-j}$ [illegible]

T[illegible] 1 [illegible] T[illegible] [illegible] M[illegible] (MARS) [illegible] (RF) [illegible] (MLR) [illegible] (GAM) [illegible] T[illegible] Table 1 [illegible] A[illegible] MARS [illegible] RF [illegible]

[illegible] RF [illegible] In [illegible] 2×10^{16}.

Fig. 4 [illegible] RF [illegible]

In [illegible] 2005 [illegible] 2006–2010. [illegible] Fig. 4 [illegible] RF [illegible] RF [illegible] MARS [illegible]

[illegible] (e.g., MLR) [illegible] GAMs [illegible] RF [illegible] MARS [illegible] MARS [illegible] (Friedman, 1991) [illegible]

[illegible] (2010), [illegible] RF [illegible] Fig. 5.

[illegible] (HDD), [illegible] U.S.

In [illegible] Fig. 5.

[illegible] (M[illegible], 2014).

C[illegible] D[illegible] (i.e., [illegible] 600) [illegible] H[illegible] i.e., HDD 3000 [illegible] N[illegible] C[illegible] U.S., [illegible] N[illegible] D[illegible], M[illegible], [illegible] (T[illegible]) T[illegible]

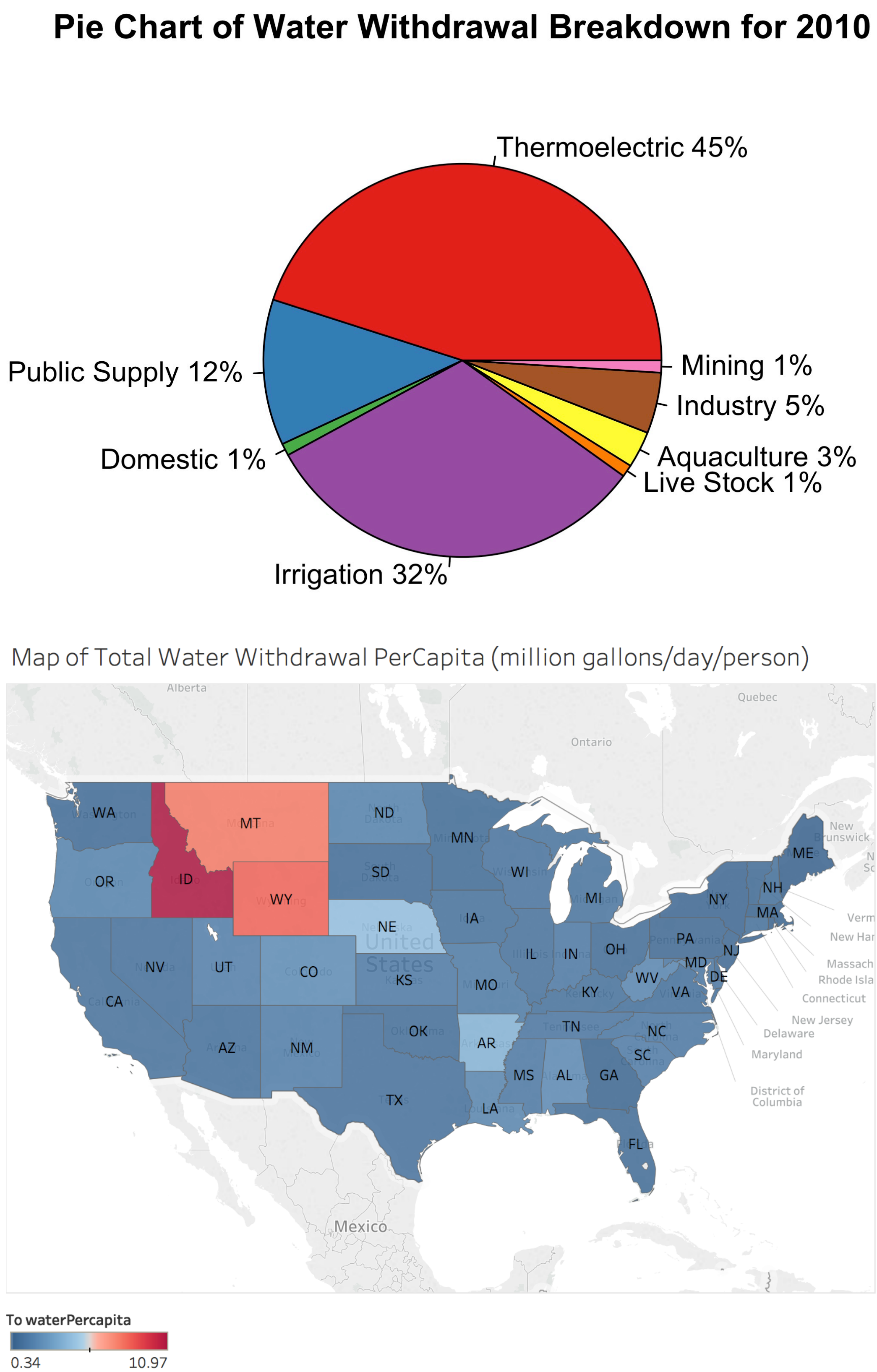
[illegible] O[illegible] U.S. ([illegible] ±10%), [illegible] ([illegible]) [illegible] ([illegible]) [illegible] RCP8.5 [illegible]. I[illegible] (e.g., HDD [illegible]) [illegible] (e.g., [illegible]); [illegible] ([illegible]) [illegible]

Summary of models performance given as correlation coefficient (R^2), fitted Root Mean Square Error (RMSE; million gal/day/person), and Leave one out cross validation (LOOCV) RMSE. Each model is trained and tested using all available data records for the period 1991-2010.

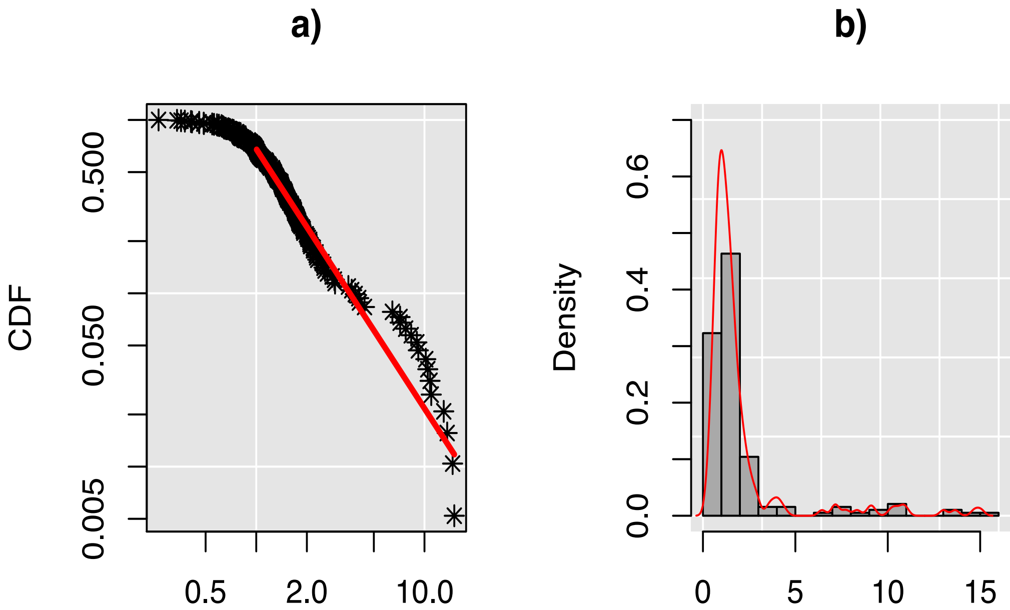
Model	R^2	RMSE	LOOCV RMSE
[illegible]	–	2.60	2.62
[illegible]	0.57	1.71	1.84
[illegible] (GAM)	0.61	1.62	1.62
[illegible]	0.85	0.99	1.40

Summary of models predictive accuracy. Each Model is trained using 1991-2005 data and tested using 2006-2010 data. Summary performance is presented here in terms of correlation coefficient (R^2), fitted Root Mean Square Error (RMSE; million gal/day/person), Leave one out cross validation (LOOCV) RMSE, and prediction RMSE (for the test data). See Appendix D for more details on LOOCV-RMSE.

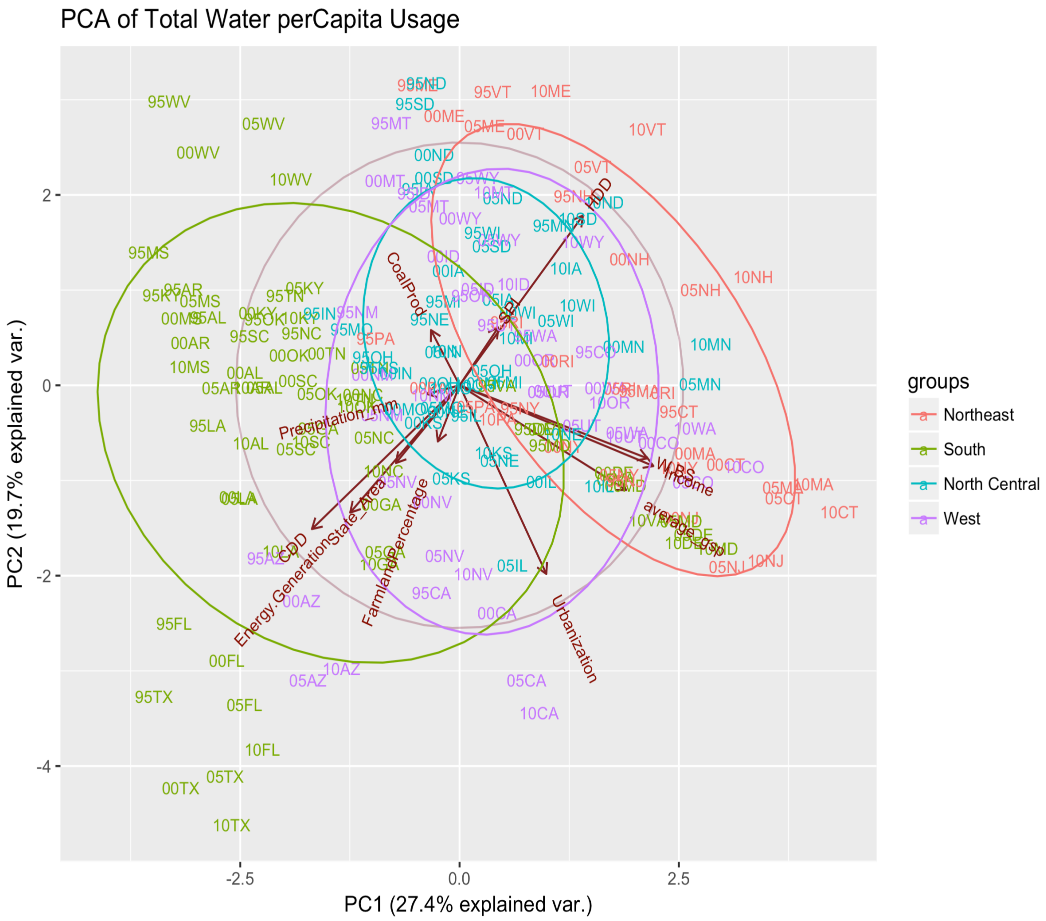
Model	R^2	RMSE	LOOCV RMSE	Prediction RMSE
[illegible]	–	2.75	2.77	2.11
[illegible]	0.59	1.76	2.00	1.52
[illegible] (GAM)	0.65	1.63	1.68	1.31
[illegible]	0.95	0.60	1.57	1.35



i Top: The breakdown of US-wide water withdrawals across the eight major sectors during the period 2006-2010. Bottom: Spatial distribution of the U.S. wide per-capita water withdrawal (in million gallons per-day).



i The empirical distribution of per-capita water withdrawals (in million gallons per day) for the period 1991-2010; (a) the red line shows that power-law fits the tail of the empirical cumulative distribution reasonably well (b) the histogram of per-capita water demand with overlain kernel density line (in red).



i Principal Component Analysis (PCA) biplot of the per-capita water usage (in million gallons per-day) for the period 19952010. The states are color-coded based on their proximity to water bodies and the two digits next to the state codes indicate the year associated with the water use data for the state.

Figure 4. Top: Scatter plot of observed versus estimated values of per-capita water withdrawal (in million gallons per-day) using data of 1995-2010. Bottom: Scatter plot of observed versus predicted values of per-capita water usage (in million gallons per-day) using data of 2006-2010. In the latter case, the models were trained using data of 1995-2005, and the testing was conducted in an independent period of 2006-2010.

RF [illegible]
SPI [illegible]
U.S. [illegible]
USD [illegible] ($)
USGS [illegible]

Acknowledgments

[illegible] NSF [illegible] EES#1555582 [illegible] C4E [illegible] P CCRC) [illegible] ISI-MIP) [illegible] CMIP, [illegible] 3. [illegible] (G [illegible]).

References

A[illegible], J., [illegible], P., H[illegible], T., K[illegible], F., L[illegible], B., [illegible], T., & S[illegible], S. (2003). [illegible] 2 [illegible], (3), 317–337.

BEA. (2017). [illegible] 2 .

B[illegible], B., & J[illegible], C. (2008). [illegible] A[illegible] D[illegible], CO.

B[illegible], L. (2001). [illegible], (1), 5–32.

C[illegible], M. K., P[illegible], L. F., & J[illegible], R. B. (2011). [illegible], (10), 6234–6242.

CSO. (2017). [illegible]

D[illegible], E. G., K[illegible], P., & E[illegible], J. A. (2013). [illegible] 2095. [illegible], 296–313.

D[illegible], E. A., M[illegible], T. A., S[illegible], R., & R[illegible], J. A. (2014). [illegible] (2), 146–159.

EIA. (2017). [illegible]

EPA. (2017). [illegible]

F[illegible], Y., & [illegible], H. (2004). [illegible] 0.5 [illegible] 1948 [illegible] (D10).

F[illegible], M., K[illegible], E., [illegible], I., E[illegible], S., [illegible], F., & A[illegible], J. (2013). [illegible] 60 [illegible] economic development: A global simulation study. Global Environmental Change, 23(1), 144{156.

Friedman, J. H. (1991). Multivariate Adaptive Regression Splines. The Annals of Statistics, 19(1), 1{67.

Genuer, R., Poggi, J.-M., & Tuleau-Malot, C. (2010). Variable selection using random forests. Pattern Recognition Letters, 31(14), 2225{2236.

Giordano, M., & Shah, T. (2014). From IWRM back to integrated water resources management. International Journal of Water Resources Development, 30(3), 364{376.

Hall, M. J., Postle, S. M., & Hooper, B. D. (1989). A data management system for demand forecasting. International Journal of Water Resources Development, 5(1), 3{10.

Hamoda, M. F. (1983). Impacts of socio-economic development on residential water demand. International Journal of Water Resources Development, 1(1), 77{84.

Hanasaki, N., Kanae, S., Oki, T., Masuda, K., Motoya, K., Shirakawa, N., ... Tanaka, K. (2008a). An integrated model for the assessment of global water resources{part 1: Model description and input meteorological forcing. Hydrology and Earth System Sciences, 12(4), 1007{1025.

Hanasaki, N., Kanae, S., Oki, T., Masuda, K., Motoya, K., Shirakawa, N., ... Tanaka, K. (2008b). An integrated model for the assessment of global water resources{part 2: Applications and assessments. Hydrology and Earth System Sciences, 12(4), 1027{1037.

Hastie, T., Tibshirani, R., & Friedman, J. (2009). Overview of Supervised Learning. Springer New York.

Hayes, M., Svoboda, M., Wall, N., & Widhalm, M. (2010). The Lincoln Declaration on Drought Indices: Universal Meteorological Drought Index Recommended. Bull. Amer. Meteor. Soc., 92(4), 485-488.

IOWA. (2017). Historical Urban Percentage of the Population for States. Source: Decennial Census, U.S. Census Bureau. Retrieved from Iowa State University, Iowa Community Indicators Program: Available at: http://www.icip.iastate.edu/tables/population/urban-pct-states. Last accessed on 04/04/2017.

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). An Introduction to Statistical Learning. Springer New York.

Johnson, B., Christopher, T., Anil, G., & NewKirk, S. V. (2011). Nebraska Irrigation Fact Sheet. Rep. no. 190. Department of Agricultural Economics, University of Nebraska Lincoln..

Jorgensen, B., Graymore, M., & O'Toole, K. (2009). Household water use behavior: An integrated model. Journal of Environmental Management, 91(1), 227{236.

Liaw, A., & Wiener, M. (2002). Classi cation and regression by randomforest. R news, 2(3), 18{22.

Lutz, J. D., Liu, X., McMahon, J. E., Dunham, C., Shown, L. J., & McCure, Q. T. (1996). Modeling patterns of hot water use in households. O ce of Scienti c and Technical Information (OSTI). Lawrence Berkeley National Laboratory, Berkeley, CA.

Maupin, M. A., Kenny, J., Hutson, S. S., Lovelace, J. K., Barber, N. L., & Linsey, K. S. (2014). Estimated use of water in the United States in 2010. US Geological Survey.

McKee, T. B., Doesken, N. J., & Kleist, J. (1993). The relationship of drought frequency and duration to time scales. Eighth Conference on Applied Climatology, Anaheim, California, 17-22 January 1993.

NOAA. (2017). National Oceanic and Atmospheric Administration (NOAA), Degree Days Statistics National Weather Service; Center for Weather and Climate Prediction. Retrieved fromm: http://www.cpc.ncep.noaa.gov/products/analysis_monitoring/cdus/degree_days/ ; Last accessed on 04/04/2017.

Pokhrel, Y. N., Hanasaki, N., Wada, Y., & Kim, H. (2016). Recent progresses in incorporating human land{water management into global land surface models toward their integration into earth system models. Wiley Interdisciplinary Reviews: Water, 3(4), 548{574.

Rahaman, M. M., & Varis, O. (2005). Integrated water resources management: evolution, prospects and future challenges. Sustainability: Science, Practice and Policy, 1(1), 15{21.

Sebri, M. (2016). Forecasting urban water demand: A meta-regression analysis. Journal of Environmental Management, 183, 777{785.

Sovacool, B. K., & Sovacool, K. E. (2009). Identifying future electricity{water tradeo s in the United States. Energy Policy, 37(7), 2763{2773.

Sutanudjaja, E. H., Beek, R. v., Wanders, N., Wada, Y., Bosmans, J. H., Drost, N., . . . others (2018). PCR-GLOBWB 2: a 5 arcmin global hydrological and water resources model. Geoscienti c Model Development, 11(6), 2429{2453.

USCB. (2017). U.S. Census Bureau, Historical Income Tables: Households. Retrieved from U.S. Census Bureau: https://www.census.gov/hhes/www/income/data/historical/household/USGS . (1995-2010).

USDA. (2007). Farm and Ranch Irrigation Survey. Retrieved from USDA, Census of Agriculture. Available from: https://www.agcensus.usda.gov/Publications/2007/Online_Highlights/Farm_and_Ranch_Irrigation_Survey/fris08.pdf ; Last accessed on 04/04/2017.

USGS. (2017). Water-use data available from USGS. Retrieved from U.S. Geological Survey http://water.usgs.gov/watuse/data/ ; Last accessed on 04/04/2017.

Wada, Y., Bierkens, M. F., Roo, A. d., Dirmeyer, P. A., Famiglietti, J. S., Hanasaki, N., . . . others (2017). Human{water interface in hydrological modelling: current status and future directions. Hydrology and Earth System Sciences, 21(8), 4169{4193.

Wada, Y., Wisser, D., & Bierkens, M. (2014). Global modeling of withdrawal, allocation and consumptive use of surface water and groundwater resources. Earth System Dynamics, 5(1), 15.

Warszawski, L., Frieler, K., Huber, V., Piontek, F., Serdeczny, O., & Schewe, J. (2014). The Inter-Sectoral Impact Model Intercomparison Project (ISI-MIP): Project framework. Proceedings of the National Academy of Sciences, 111(9), 3228{3232.