Supplemental material: Predicting Inundation Dynamics and Hydroperiods of Small, Isolated Wetlands Using a Machine Learning Approach
Journal: *Wetlands*

Jeffrey W. Riley* U.S. Geological Survey, South Atlantic Water Science Center (ORCID 0000-0001-5525-3134)

Charles C. Stillwell U.S. Geological Survey, South Atlantic Water Science Center (ORCID 0000-0002-4571-4897)

*Corresponding Author: jriley@usgs.gov

Table S1. Characteristics of each wetland in the study. In the cases where the notes column indicates that the wetland was not in the national wetlands inventory (NWI), the NWI_classification (based on Cowardin et al 1979) was determined based on the classification of similar wetlands or identified as a sinkhole wetland.

wetland id	**area hectares**	**dominant surrounding vegetation**	**NWI classification**	**notes**
30knot	0.19	conifer	freshwater emergent wetland	broad shallow basin with deeper sinkhole on one end that holds water permanently
biggins	1.38	riparian	freshwater pond	wet prairie with possible sinkhole on one end of basin
borrow405	0.41	shrubland	freshwater pond	man-made borrow pit that looks and functions like adjacent natural depressions
burnt	0.16	conifer	freshwater forested/shrub wetland	not in NWI dataset
chicky	0.20	conifer	freshwater forested/shrub wetland	not in NWI dataset

cnb	0.24	conifer	freshwater emergent wetland	not in NWI dataset
corner	0.35	conifer	freshwater pond	Not applicable (NA)
dh	0.13	conifer	sinkhole	not in NWI dataset, sinkhole wetland in sandhill uplands
en	0.17	hardwood	freshwater forested/shrub wetland	not in NWI dataset
fn	0.32	conifer	freshwater forested/shrub wetland	NA
fwc	0.35	conifer	freshwater forested/shrub wetland	NA
hawk	0.16	conifer	freshwater forested/shrub wetland	NA
itw	0.25	conifer	freshwater forested/shrub wetland	NA
jbo	0.16	conifer	sinkhole	not in NWI dataset, deep forested basin in upland sandhills
jen	0.01	conifer	sinkhole	not in NWI dataset, small, deep basin in upland sandhills
jr10	0.11	conifer	freshwater forested/shrub wetland	not in NWI dataset
jr18	0.26	conifer	freshwater pond	NA
jr36	0.46	riparian	freshwater forested/shrub wetland	side lobe of larger wetland complex
kf	0.23	conifer	freshwater pond	NA

pond10	0.62	riparian	freshwater pond	NA
pond2	0.24	conifer	freshwater emergent wetland	NA
pond7	0.39	conifer	freshwater forested/shrub wetland	broad shallow basin with sinkhole on one end that holds water permanently
pond8	0.21	conifer	freshwater forested/shrub wetland	NA
ptess	0.28	conifer	freshwater forested/shrub wetland	NA
pygmy	0.07	conifer	freshwater emergent wetland	not in NWI dataset
rnd113	0.28	conifer	freshwater forested/shrub wetland	not in NWI dataset
rnd117	0.21	conifer	freshwater forested/shrub wetland	not in NWI dataset
rnd120	0.26	conifer	freshwater forested/shrub wetland	NA
rnd123	0.27	conifer	freshwater forested/shrub wetland	NA
rnd125	1.06	riparian	freshwater pond	side lobe of large swamp
rnd42	0.31	conifer	freshwater emergent wetland	NA
rnd49	0.21	conifer	freshwater emergent wetland	NA

rnd50	0.24	conifer	freshwater pond	NA
rnd52	0.27	conifer	freshwater emergent wetland	NA
rnd62	0.18	conifer	freshwater pond	NA
rnd70	0.29	riparian	freshwater pond	NA
rnd72	0.26	conifer	freshwater emergent wetland	NA
rnd75	0.21	riparian	freshwater emergent wetland	NA
rnd76	0.26	hardwood	freshwater pond	NA
rnd80	0.23	conifer	freshwater pond	NA
rnd81	0.46	riparian	freshwater emergent wetland	NA
rnd82	0.17	conifer	freshwater pond	NA
rnd83	0.55	riparian	freshwater pond	NA
sb	0.12	conifer	freshwater forested/shrub wetland	not in NWI dataset
sedge	0.26	conifer	freshwater pond	NA
taint	0.28	conifer	freshwater forested/shrub wetland	NA
talp	0.26	conifer	freshwater emergent wetland	NA
w1	0.30	conifer	freshwater forested/shrub wetland	NA

w2	0.39	conifer	freshwater forested/shrub wetland	NA
w3	0.40	conifer	freshwater forested/shrub wetland	NA
w4	0.51	riparian	freshwater pond	NA
wbf	0.15	conifer	freshwater forested/shrub wetland	not in NWI dataset
wpt150	0.10	conifer	freshwater forested/shrub wetland	not in NWI dataset
wpt19	0.18	conifer	freshwater emergent wetland	NA
wpt192	0.15	conifer	freshwater pond	not in NWI dataset
wpt57	0.29	conifer	freshwater emergent wetland	NA
wpt68	0.18	conifer	freshwater forested/shrub wetland	NA
wpt69	0.18	conifer	freshwater emergent wetland	NA
wpt79	0.20	conifer	freshwater forested/shrub wetland	NA

S2. Determining the best thresholds for splitting the data

We undertook a rigorous multi-faceted approach to model evaluation that tested performance on unseen wetlands and on unseen observations from wetlands in the training set. To evaluate the best strategy for splitting the dataset we explored many options (table S2). We first split the data into training and validation groups using a stratified Monte Carlo cross validation approach (Kuhn and Johnson 2018), where all data from 3-12 randomly selected wetlands (equal numbers from permanent, frequently inundated, and infrequently inundated wetlands) were set aside as the "validation set." The remaining data were randomly split into 60/40, 70/30, 80/20, 90/10, and 95/5 train/test sets to determine which split resulted in the best performance. To evaluate the model performance on the different data splits we relied on two metrics that are useful for evaluating model performance on imbalanced datasets (balanced accuracy and sensitivity). Balanced accuracy quantifies the model's ability to predict both classes and sensitivity quantifies the model's ability to predict the focal class – in this case – dry days, (the minority classification). For each one of the data splits, 50 model simulations were conducted and summary statistics from those were used to evaluate the best splitting strategy (table S2). Based on the metrics we evaluated, there was not a clear "best" splitting strategy, and several combinations provided reasonable results (Table S2). The out-of-bag (OOB) error rate and the balanced accuracy on the test set decreased (albeit very slightly) as the proportion of training data increased. However, this was not the case for balanced accuracy or sensitivity on the validation set where these metrics varied more randomly and were generally lower as the number of validation wetlands exceeded six. Therefore, in the manuscript we present results from models with six validation wetlands and 70/30 train/test split. This data configuration presents a balance of good performance and allows an evaluation of prediction on six unseen wetlands for each model iteration.

Table S2. Evaluation of model performance based on different data splitting scenarios. Bold text indicates the data splits that were used in the manuscript.

Number of validation wetlands withheld	Train/test split	OOB[a] median	Test set[b] balanced accuracy median	Test set balanced accuracy MAD[c]	Validation[d] set balanced accuracy median	Validation set balanced accuracy MAD	Validation set sensitivity median	Validation set sensitivity MAD
3	60/40	0.060	0.961	0.001	0.899	0.045	0.888	0.096
3	70/30	0.056	0.964	0.001	0.870	0.095	0.856	0.197
3	80/20	0.052	0.967	0.002	0.837	0.124	0.756	0.280
3	90/10	0.050	0.969	0.002	0.832	0.124	0.770	0.262
3	95/5	0.048	0.970	0.002	0.802	0.095	0.785	0.239
6	60/40	0.060	0.961	0.001	0.844	0.073	0.848	0.132
6	**70/30**	**0.055**	**0.964**	**0.001**	**0.875**	**0.067**	**0.844**	**0.170**
6	80/20	0.052	0.966	0.002	0.845	0.075	0.848	0.145
6	90/10	0.050	0.968	0.002	0.818	0.087	0.746	0.196
6	95/5	0.048	0.969	0.003	0.830	0.067	0.809	0.169
9	60/40	0.060	0.962	0.002	0.823	0.071	0.772	0.189
9	70/30	0.056	0.965	0.001	0.788	0.101	0.725	0.193
9	80/20	0.053	0.967	0.002	0.831	0.090	0.803	0.169
9	90/10	0.050	0.968	0.002	0.813	0.069	0.745	0.145
9	95/5	0.049	0.969	0.002	0.835	0.068	0.783	0.148
12	60/40	0.059	0.962	0.002	0.825	0.058	0.757	0.114
12	70/30	0.055	0.964	0.001	0.810	0.095	0.719	0.198
12	80/20	0.053	0.966	0.002	0.813	0.076	0.728	0.176
12	90/10	0.049	0.968	0.002	0.801	0.068	0.691	0.161
12	95/5	0.049	0.970	0.003	0.802	0.076	0.702	0.204

a. Out of bag error rate (OOB) is based on prediction of those observations that weren't included in the bootstrap sample.
b. Test set refers to the observations that were held out for testing from wetlands that were included in the training set. No further metrics are included from the test data because the balanced accuracy is high and stable across all split scenarios indicating the models are performing well when they have "seen" data from a given wetland regardless of how much is included (within the ranges evaluated).
c. MAD is the median absolute deviation for each metric for the 50 model iterations.
d. Validation set refers to the withheld wetlands and was used to test the model's ability to predict in unseen wetlands.

Disclaimer

Any use of trade, firm, or product names is for descriptive purposes only and does not imply endorsement by the U.S. Government.

References

Cowardin LM, Carter V, Golet FC, LaRoe ET (1979) Classification of wetlands and deepwater habitats of the United States. Washington, DC

Kuhn, M., & Johnson, K. (2018). Applied predictive modeling. Springer.