Boreal Wetland probability - technical documentation

ABMI Geospatial Centre

December, 2017

Contents

1 Introduction

In Alberta it is estimated that about 20% of the province (66,000 km^2) is covered by wetlands of which 90% are in the boreal forest (Alberta Environment and Sustainable Resource Development, 2013). Much of the native wetland areas in Alberta have already been drained or permanently modified and other areas are currently undergoing rapid changes due to resource exploration and extraction (Ducks Unlimited, 2017). For this reason it is increasingly important to have spatially extensive maps of wetland type and extent.

Currently, the best source for wetland inventory in Alberta is the Alberta Merged Wetland Inventory (AMWI) (Alberta Environment and Parks, 2012). This represents an amalgamated data source from 33 different data sources with varying methodologies. Other products such as Alberta Vegetation Inventory Enhanced (AVIE) (Alberta Environment and Parks, formerly ESRD, 2016), Primary Land and Vegetation Inventory (PLVI) (Alberta Environment and Parks, formerly ESRD, 2012), and Derived Ecosite Phase (DEP) (Alberta Agriculture and Forestry, 2017) provide quality information on wetland extent and type but do not provide Alberta-wide spatial coverage. The lack of spatially consistent and extensive wetland maps becomes a problem for various monitoring, and policy frameworks across Alberta such as the Biodiversity Management Framework (BMF). In the Lower Athabasca BMF Aquatic Habitat and Fen Cover are two indicators requiring wetland data. The varying sources of AMWI and incomplete spatial coverage of AVIE, PLVI and DEP could result in spatially inconsistent results. Examples like these demonstrate the need for a consistent spatially extensive dataset for wetland and wetland types.

Satellite imagery and other remotely sensed data offers means to generate a spatially extensive and consistent dataset. Within the boreal, AMWI used a either just optical imagery or a mix of optical remote sensing, and Synthetic Aperture Radar (SAR) data (Ducks Unlimited Canada, 2011) along with ancillary data such as a digital elevation model (DEM), forest inventories, and fire information for their wetland mapping. Recent literature suggests that a combined approach of optical, SAR, and high resolution DEMs may be the most effective for wetland mapping (Touzi *et al.,* 2011; Brisco, 2015; Difebo *et al*., 2015; Hird *et al.,* 2017). At this time, the best freely accessible source for this may be Sentinel-1 and-2 data (Copernicus [2014, 2015, 2016]), which offers decent spatial and temporal resolutions (10-m resolution, and 5 – 6 day revisit time) for mapping wetlands on regional to provincial scales. Combining these datasets with a fine resolution DEM should provide good information for classifying wetlands with remote sensing data.

To summarize, there are two goals for this project:

1. To develop a framework in which wetland occurrence can be predicted across large areas, at regular intervals, with easily accessible/open source input data and software.
2. To generate the most up-to-date and accurate data set of wetland occurrence for the entire boreal region.

2 Methods

2.1 Study area

The study area primarily consists of the Boreal Natural Region of Alberta with small parts of the foothills, parkland, and Canadian Shield included. This study area makes up about 60% (397, 958 km^2) of the total area of Alberta.

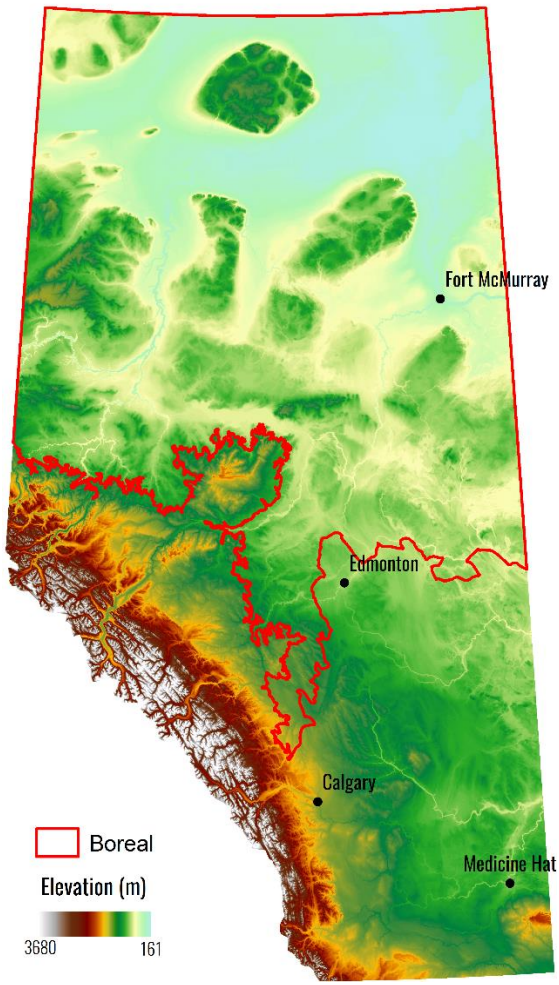


Figure 1: The spatial delineation of the boreal region.

2.2 Data

Sentinel-1, -2 (Copernicus [2016, 2017]), LiDAR Digital Terrain Model (DTM) (Government of Alberta, 2006), and Shuttle Radar Topography Mission (SRTM) Digital Elevation Model (DEM) (USGS, 2006) data was used to generate wetland probability in the boreal region. All Sentinel and SRTM data were acquired, processed, and downloaded through Google Earth Engine (GEE) (Google Earth Engine Team, 2015). GEE stores Sentinel-1 (SAR imagery) ground range detected scenes which have been pre-processed with the Sentinel-1 Toolbox (Sentinel Application Platform – Sentinel-1 Toolbox). These pre-processing steps include thermal noise removal, radiometric calibration, and terrain correction (Google Earth Engine Team, 2015). Dual polarization (VV VH) Sentinel-1 (S1) images were further processed in the GEE environment by performing an incidence angle correction (Gauthier *et al.*, 1998) and smoothing with a 3x3 Sigma Lee filter (Lee *et al.*, 2009) (credit to Guido Lemoine for GEE code). Once all S1 images were processed, a normalized difference of polarization (NDPOL) was calculated (see Table 1) and added to the available bands. To generate a single composite image for the S1 variables the per pixel mean of the VH and NDPOL bands were calculated. A total of 478 S1 images were used in the calculation of the VH and NDPOL variables.

Sentinel-2 (optical imagery) top of atmosphere data was acquired through GEE. Clouds, shadows, snow, and ice were removed with the QA60 band (a quality control band used to identify bad pixels) and further

cloud masking was done using bands 1 (aerosols) and 11 (cloud). Sentinel-2 (S2) images intersecting with the boreal region during 2016-2017 leaf-on season (May 15 – August 31) were used to generate vegetation indices, and Principal component 1 and 2. PC1/2 metrics were generated with the 10m S2 bands (B2, B3, B4, and B8) in a principal component analysis. The merged S2 data was generated using a median composting algorithm where the median time series value for each pixel was selected as the most representative pixel. A total of 3,148 S2 images were used in the calculation of the vegetation indices and PC bands.

The DEM/DTM data used for modelling came from three sources: 1m bare earth (BE) LiDAR for the forest regions of Alberta(Government of Alberta, 2006), 15m bare earth (BE) LiDAR from the prairie regions of Alberta(Government of Alberta, 2017), and 30m SRTM DEM data for anywhere without LiDAR (USGSm 2006). The 1m BE LiDAR was mean aggregated to 10m to match the S1 and S2 data and the 15m BE LiDAR was resampled to 10m using cubic convolution method. The SRTM data was turned into a floating point raster, then resampled to 10m resolution using cubic convolution and then subsequently smoothed using a 7x7 pixel mean filter. Two topographic indices (TWI and TPI, Table 1) were calculated separately for each DEM data set and then merged when complete. All topographic indices were calculated in SAGA version 5.0.0 (Conrad *et al.*, 2015). All the input variables can be seen in Figure 2 and the equations and description can be seen in Table 1.

Training data was taken from the Alberta Biodiversity Monitoring Institute 3x7km Land Cover Photoplots (hereafter 3x7s) (ABMI, 2016). These photoplots are derived from high resolution 3D image interpretation and give detailed attribution of land cover information. They are typically very accurate with less than 1% of features possessing errors (ABMI, 2016).

Table 1: List of possible input variables used in the wetland probability model

Variable	Data source	Equation	Description
ARI	Sentinel-2	$\left(\frac{Band\ 8}{Band\ 2}\right) - \left(\frac{Band\ 8}{Band\ 3}\right)$	Anthocyanin Reflectance Index. An index sensitive to anthocyanin pigments in plant foliage (Gitelson *et al.*, 2001).
NDVI	Sentinel-2	$\frac{(Band\ 8 - Band\ 4)}{(Band\ 8 + Band\ 4)}$	Normalized Difference Vegetation Index. Index for estimating photosynthetic activity, and leaf area (Rouse *et al.,* 1973).
NDWI	Sentinel-2	$\frac{(Band\ 3 - Band\ 8)}{(Band\ 3 + Band\ 8)}$	Normalized difference Water Index from Mcfeeters (1996)
NDPOL	Sentinel-1	$\frac{(VH - VV)}{(VH + VV)}$	Normalized Difference of Polarization.
PC1	Sentinel-2	-	The first principal component of variation of Bands 2, 3, 4, and 8 of Sentinel-2 data
PC2	Sentinel-2	-	The second principal component of variation of Bands 2, 3, 4, and 8 of Sentinel-2 data
PSRI	Sentinel-2	$\frac{(Band\ 4 - Band\ 2)}{(Band\ 5)}$	Plant Senescence Reflectance Index. A ratio used to estimate the ratio of bulk carotenoids to chlorophyll (Hatfield and Prueger, 2010).
REIP	Sentinel-2	$702 + 40\left(\frac{\left(\frac{Band\ 4 + Band\ 7}{2}\right) - Band\ 5}{(Band\ 6 - Band\ 5)}\right)$	Red Edge Inflection Point. An approximation on a hyperspectral index for estimating the position (in nm) of the NIR/red inflection point in vegetation spectra (Herrmann, *et al.,* 2011).

TPI	LiDAR, SRTM DEMs	-	Topographic Position Index (TPI) generated in SAGA (Conrad *et al.*, 2015). An index describing the relative position of a pixel within a valley, ridge top continuum calculated in a given window size. TPI was calculated with a 500m moving window for this purpose (Weiss, 2001).
TWI	LiDAR, SRTM DEMs	-	Saga Wetness Index. A SAGA (Conrad *et al.*, 2015) version of the Topographic Wetness Index. Potential wetness of the ground based on topography (Böhner *et al.*, 2002).
VH	Sentinel-1	-	Vertical polarization sending horizontal polarization receiving SAR backscatter in decibels.

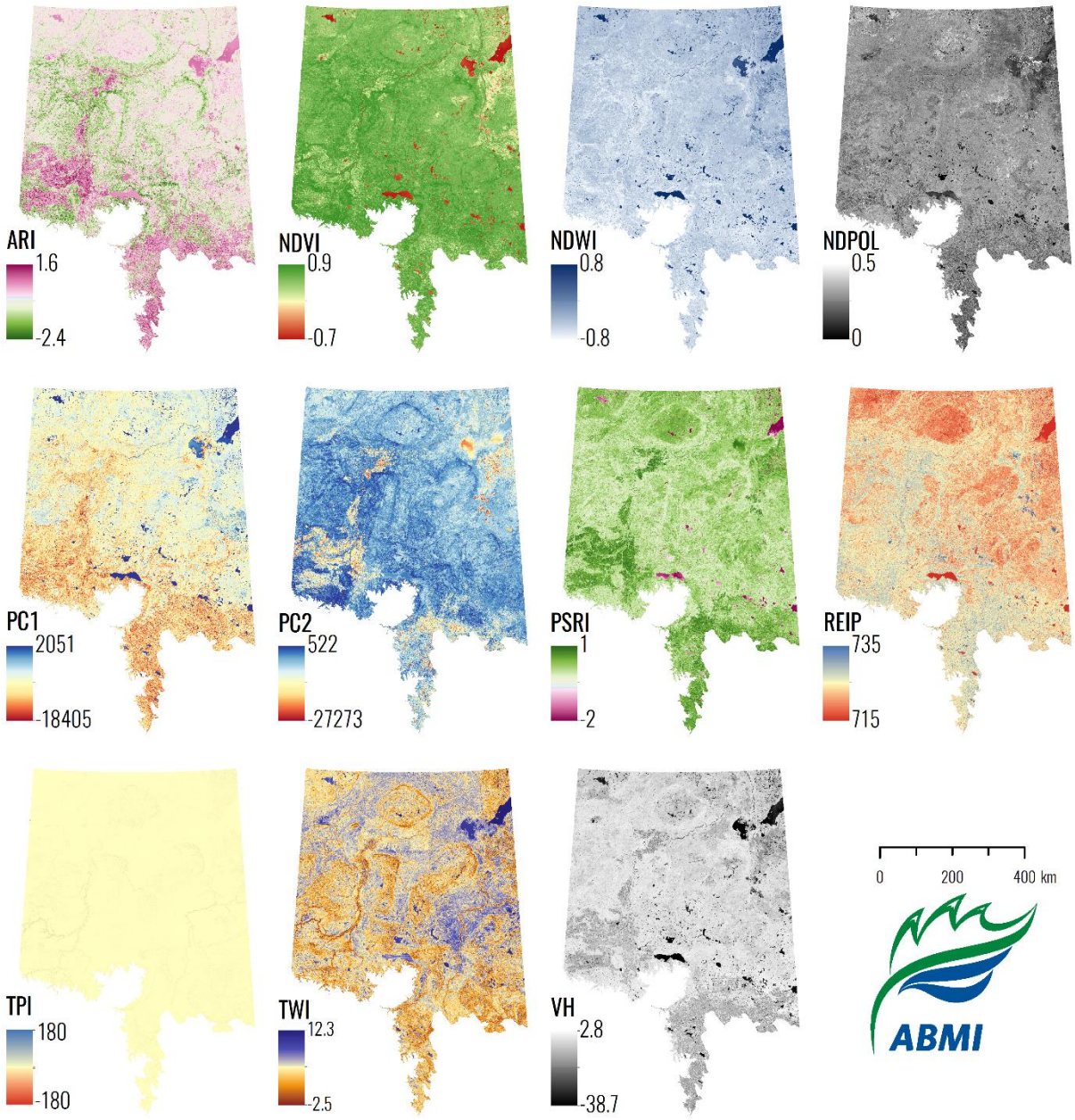
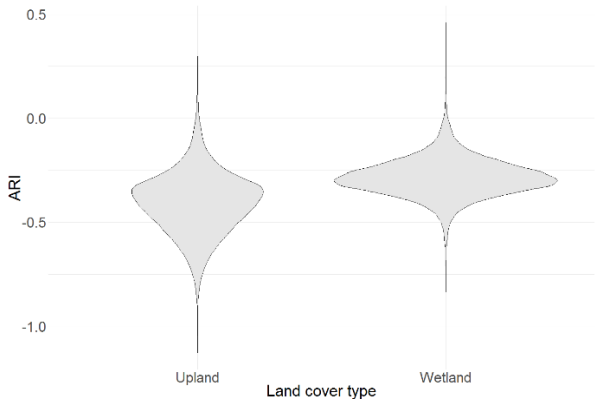
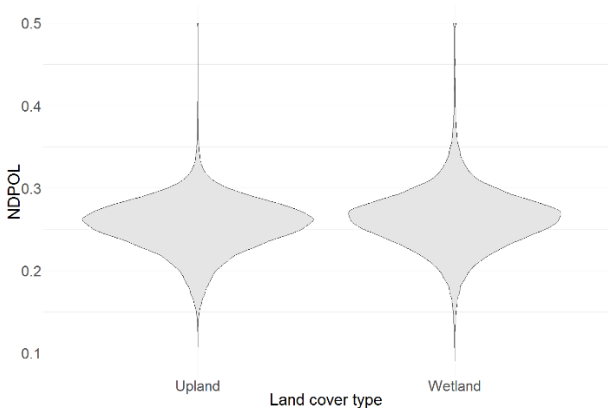


Figure 2: The 11 potential input raster inputs for the wetland probability model. NDPOL and VH are generated from Sentinel-1, TPI and TWI are generated from DEM data, and the rest are generated from Sentinel-2 data.

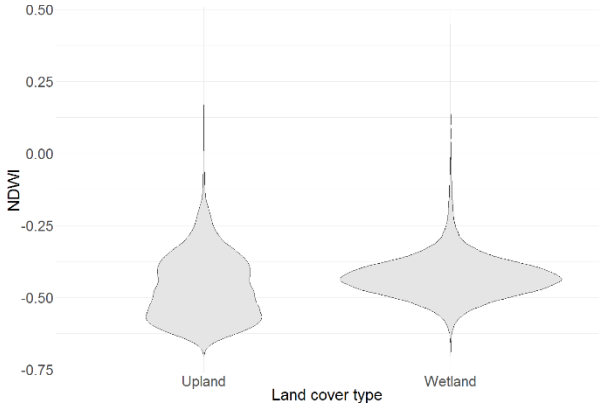
2.3 Data exploration and variable selection

Figure 3 shows the distribution of values for all possible input variables divided by wetland and upland class. The wetland and upland training data comes from 3x7s. Many variables show a distinct difference between wetlands and uplands (ARI, NDVI, PC2, REIP, TWI, and VH). Some variables show similar values between classes but demonstrate different distributions around the same mean (NDPOL, NDWI, PC1, and TPI).


ARI
0.5
0.0
-0.5
-1.0
Upland
Wetland
Land cover type

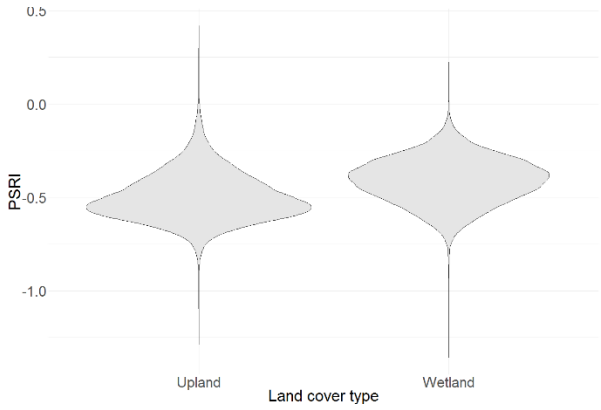

NDPOL
0.5
0.4
0.3
0.2
0.1
Upland
Wetland
Land cover type

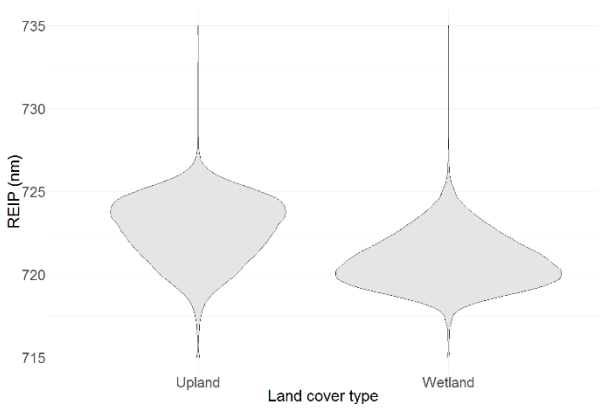

NDVI
0.6
0.3
0.0
-0.3
Upland
Wetland
Land cover type


NDWI
0.50
0.25
0.00
-0.25
-0.50
-0.75
Upland
Wetland
Land cover type


PC1
2000
1000
0
-1000
-2000
Upland
Wetland
Land cover type


PC2
500
250
0
-250
-500
Upland
Wetland
Land cover type


PSRI
0.5
0.0
-0.5
-1.0
Upland
Wetland
Land cover type


REIP (nm)
735
730
725
720
715
Upland
Wetland
Land cover type

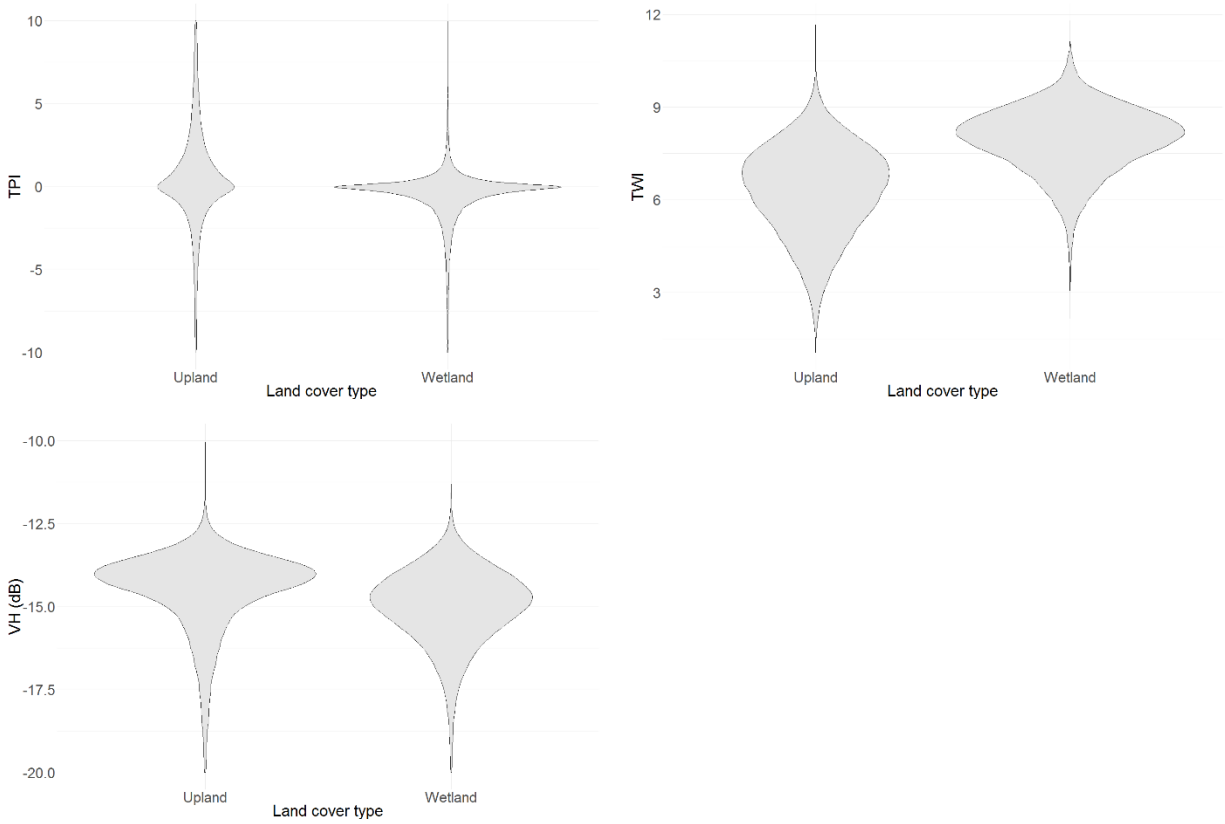


Figure 3: Distribution all possible input variables divided by upland or wetland class.

Using information from Figure 3, Table 2, and Table 3 we selected the most important variables for predictive modelling of wetlands. TWI, ARI, and REIP all have good separation between wetland/upland classes, high correlation to a binary wetland/upland layer, and high relative importance in a predictive model. TPI and NDPOL are chosen due to their low correlation to TWI, ARI, and REIP, and since they show moderate importance in the predictive model. The final chosen variable is NDWI due to moderate correlation with wetlands, and lower correlation with all other variables. The highest cross correlation between the six chosen variables is (-0.61) between ARI and REIP which was deemed acceptable. The chosen variables for predictive modelling are highlighted in green in Table 3. These six variables provide a good mix of data sources (three from S2, two from DEMs, and one from S1).

Table 2: Cross correlation between the 11 possible input variables and a binary wetland/upland grid.

	ARI	NDPOL	NDVI	NDWI	PC1	PC2	PSRI	REIP	TPI	TWI	VH	wetland
ARI	1.00	-0.11	-0.43	0.35	0.02	-0.68	0.74	-0.61	-0.07	0.33	-0.48	0.44
NDPOL	-0.11	1.00	0.18	-0.15	-0.04	0.24	-0.13	-0.03	-0.05	0.06	0.18	0.11
NDVI	-0.43	0.18	1.00	-0.94	-0.72	0.72	-0.33	0.44	0.04	-0.15	0.58	-0.20
NDWI	0.35	-0.15	-0.94	1.00	0.87	-0.55	0.09	-0.46	-0.05	0.15	-0.43	0.20
PC1	0.02	-0.04	-0.72	0.87	1.00	-0.19	-0.26	-0.36	-0.05	0.09	-0.13	0.12
PC2	-0.68	0.24	0.72	-0.55	-0.19	1.00	-0.77	0.43	0.03	-0.20	0.65	-0.26
PSRI	0.74	-0.13	-0.33	0.09	-0.26	-0.77	1.00	-0.46	-0.02	0.23	-0.58	0.33

REIP	-0.61	-0.03	0.44	-0.46	-0.36	0.43	-0.46	1.00	0.03	-0.34	0.35	-0.47
TPI	-0.07	-0.05	0.04	-0.05	-0.05	0.03	-0.02	0.03	1.00	-0.07	0.02	-0.12
TWI	0.33	0.06	-0.15	0.15	0.09	-0.20	0.23	-0.34	-0.07	1.00	-0.15	0.55
VH	-0.48	0.18	0.58	-0.43	-0.13	0.65	-0.58	0.35	0.02	-0.15	1.00	-0.20
wetland	0.44	0.11	-0.20	0.20	0.12	-0.26	0.33	-0.47	-0.12	0.55	-0.20	1.00

Table 3: The relative importance of the 11 possible input variables in the boosted regression tree wetland predictive model. Variables chosen for predictive modelling are seen in green.

Variable	Relative importance
TWI	38.92
ARI	16.14
REIP	10.20
TPI	9.39
NDVI	5.26
NDPOL	5.03
NDWI	4.00
VH	3.55
PC1	3.01
PC2	2.68
PSRI	1.83

2.4 Wetland classification – machine learning algorithm

To classify the probability of wetland occurrence a machine learning algorithm was developed in R Statistical Software (R Core Team, 2013). This algorithm uses a boosted regression tree modelling approach (Elith *et al.,* 2008). To build a model 1,900 random points were placed at a distance of at least 1 kilometer apart in known wetland and upland areas delineated by the 3x7s. Training points were not placed in any locations within known human footprint features or areas with open water. The spatial delineations of these features are taken from the ABMI's Human Footprint Inventory (ABMI, 2017) and the ABMI's Boreal surface water inventory (ABMI, 2017). The training wetland/upland data was taken from the 3x7s. From these 1,900 points a data frame was built describing the values of the six input variables and their corresponding binary wetland/upland information. This data frame was then put into the boosted regression tree modelling function using a tree complexity of 5, learning rate of 0.005, and bag fraction of 0.5 (Hird *et al.,* 2017). This model output: responses for the four input variables, variables importance, and Area Under the Receiver Operating Characteristic Curve (AUROC) value. The model was then used to predict wetland probability given the six input variables. This process was repeated 40 times which generated 40 wetland probability grids. This was done to reduce statistical overfitting and spatial auto-correlation (Parisien *et al.,* 2011). The mean value of these 40 grids was used to produce the final wetland probability grid. Wetlands were then classified as any value above of a probability threshold of 0.35 (see Figure 7) resulting in a binary wetland(1)/upland(0) raster. The whole R script can be seen in Script 1.

```
#################################################################
#---------------------------------------------------------------
#---------------------------------------------------------------
```

```r
#Boreal wetland probability
#Filename: "Boreal_WetlandProbability.R"
#Written and developed by Evan R. DeLancey - GIS Land Use Analyst
#Alberta Biodiversity Monitoring Institute, Nov, 14, 2017
#---------------------------------------------------------------
#---------------------------------------------------------------
######################################################################

#Load libraries
library(raster)
library(rgdal)
library(ggplot2)
library(dplyr)
library(caret)
library(snow)
library(rgeos)
library(RPyGeo)
library(dismo)
library(gbm)
library(ggthemes)

tt1 <- Sys.time()

#location of input rasters
location <- "J:/LandCover/ProbabilityofWetArea/Boreal/processed"
#location of land and wetland shapefiles
location.wetlandLand <- "J:/LandCover/ProbabilityofWetArea/Boreal/training"
#location of outputs
outputs <- "J:/LandCover/ProbabilityofWetArea/Boreal/OUTPUTS"
#version of the outputs
OutNum <- "_v1"
#enter number of iteration of subsampling
iter <- 40
#set your python location
py.loc <- "C:/Python27/ArcGIS10.4"
#Set minimum distance of sample wetland points
MinDist <- 1000

#set location of a temporary raster dump
#this can take up 100-300BG per run but is deleted after
rasterOptions(maxmemory = 1e+09,tmpdir = "J:/RtmpRasterDump")



#DEFINE FUNCTIONS
######################################################################
#---------------------------------------------------------------
#---------------------------------------------------------------

#1) set up function to get random points give location of your land and wetland shapefile and the minimum distance
pts.gen <- function(directory, mindist){
  #set up ArcGIS environment
  w <- directory
  env <- rpygeo.build.env(python.path = py.loc, workspace = directory, overwrite=1)
  rpygeo.geoprocessor("CreateRandomPoints_management", c("", "pts.shp", constraining_feature_class = "Boreal.shp",
    constraining_extent = "", number_of_points_or_field = 80000, minimum_allowed_distance = mindist), env = env, detect.required.extensions = T)
  pts <- readOGR(w, "pts")
  return(pts)
}

#2)multiplot function
multiplot <- function(..., plotlist=NULL, file, cols=1, layout=NULL) {
 library(grid)

 # Make a list from the ... arguments and plotlist
 plots <- c(list(...), plotlist)

 numPlots = length(plots)

 # If layout is NULL, then use 'cols' to determine layout
 if (is.null(layout)) {
  # Make the panel
  # ncol: Number of columns of plots
  # nrow: Number of rows needed, calculated from # of cols
  layout <- matrix(seq(1, cols * ceiling(numPlots/cols)),
                   ncol = cols, nrow = ceiling(numPlots/cols))
 }

 if (numPlots==1) {
  print(plots[[1]])

 } else {
  # Set up the page
```

```
    grid.newpage()
    pushViewport(viewport(layout = grid.layout(nrow(layout), ncol(layout))))

    # Make each plot, in the correct location
    for (i in 1:numPlots) {
      # Get the i,j matrix positions of the regions that contain this subplot
      matchidx <- as.data.frame(which(layout == i, arr.ind = TRUE))

      print(plots[[i]], vp = viewport(layout.pos.row = matchidx$row,
                                      layout.pos.col = matchidx$col))
    }
  }
}
#------------------------------------------------------------
#------------------------------------------------------------
#################################################################



#Name vars and get min and max for response curves
#################################################################
#------------------------------------------------------------
#------------------------------------------------------------
#set input varibles
wetland.tifs <- c("ARI.tif", "NDPOL.tif", "NDWI.tif", "REIP.tif", "TPI.tif", "TWI.tif")
wetland.colnames <- c("ARI", "NDPOL", "NDWI", "REIP", "TPI", "TWI", "wetland")
wetland.bricknames <- c("ARI", "NDPOL", "NDWI", "REIP", "TPI", "TWI")
fls <- wetland.tifs

setwd(location)
#extract min and max of all input rasters for response curves
r <- raster(fls[1])
min1 <- cellStats(r, 'min')
max1 <- cellStats(r, 'max')
r <- raster(fls[2])
min2 <- cellStats(r, 'min')
max2 <- cellStats(r, 'max')
r <- raster(fls[3])
min3 <- cellStats(r, 'min')
max3 <- cellStats(r, 'max')
r <- raster(fls[4])
min4 <- cellStats(r, 'min')
max4 <- cellStats(r, 'max')
r <- raster(fls[5])
min5 <- cellStats(r, 'min')
max5 <- cellStats(r, 'max')
r <- raster(fls[6])
min6 <- cellStats(r, 'min')
max6 <- cellStats(r, 'max')
minval <- rbind(min1, min2, min3, min4, min5, min6)
maxval <- rbind(max1, max2, max3, max4, max5, max6)
mm.df <- data.frame(wetland.bricknames, minval, maxval)

#binary training raster

wetland <- raster("J:/LandCover/ProbabilityofWetArea/Boreal/training/wetland.tif")
#------------------------------------------------------------
#------------------------------------------------------------
#################################################################



#BUILD MODEL 1
#####################################################################
#------------------------------------------------------------
#------------------------------------------------------------
#define fit AUC and dev
AUC <- vector()
dev <- vector()
pts.model <- pts.gen(location.wetlandLand, MinDist)
dat <- data.frame(row=1:length(pts.model))
for (i in 1:length(fls)){
  r <- raster(fls[i])
  ext <- extract(r,pts.model)
  dat <- cbind(dat,ext)
  print(paste0("done ",fls[i]))
}
ext <- extract(wetland, pts.model)
dat <- cbind(dat,ext)
dat <- dat[,-1]
colnames(dat) <- wetland.colnames
dat <- na.omit(dat)
cor(dat)
```

```
#defint model list
fit <- list()
#build model
fit[[1]] <- gbm.step(dat, 1:length(fls), length(fls)+1, family = "bernoulli", tree.complexity = 5,
   learning.rate = 0.005, bag.fraction = 0.5, silent = TRUE, warnings = FALSE)
df.importance <- data.frame(summary(fit[[1]]))
df.importance <- arrange(df.importance, var)
df.importance <- df.importance[,2]
#model stats
AUC[1] <- fit[[1]]$cv.statistics$discrimination.mean
dev[1] <- (fit[[1]]$self.statistics$mean.null - fit$self.statistics$mean.resid) / fit$self.statistics$mean.null

response.df <- data.frame(dummy=c(1:1001))
for (n in wetland.bricknames){
   d <- plot.gbm(fit[[1]], i.var = n, return.grid=TRUE, type="response")
   get.min.max <- filter(mm.df, wetland.bricknames == n)
   mn <- get.min.max[,2]
   mx <- get.min.max[,3]
   xout <- seq(mn, mx, (mx-mn)/1000)
   d <- approx(d[,1], d[,2], xout = xout, rule=2)
   d <- as.data.frame(d)
   response.df <- cbind(response.df,d)
}
#-----------------------------------------------------
#-----------------------------------------------------
#####################################################################
```

```
#BUILD MODEL ALL MODELS AND OUTPUT STATS
#####################################################################
#-----------------------------------------------------
#-----------------------------------------------------
setwd(location)


for (i in 2:iter){
   pts.model <- pts.gen(location.wetlandLand, MinDist)
   dat <- data.frame(row=1:length(pts.model))
   for (j in 1:length(fls)){
      r <- raster(fls[j])
      ext <- extract(r,pts.model)
      dat <- cbind(dat,ext)
   }
   ext <- extract(wetland, pts.model)
   dat <- cbind(dat,ext)
   dat <- dat[,-1]
   colnames(dat) <- wetland.colnames
   dat <- na.omit(dat)

   #build model
   fit[[i]] <- gbm.step(dat, 1:length(fls), length(fls)+1, family = "bernoulli", tree.complexity = 5,
      learning.rate = 0.005, bag.fraction = 0.5, silent = TRUE, warnings = FALSE)
   v.importance <- as.data.frame(summary(fit[[i]]))
   v.importance <- arrange(v.importance, var)
   v.importance <- v.importance[,2]
   df.importance <- cbind(df.importance, v.importance)

   for (n in wetland.bricknames){
      d <- plot.gbm(fit[[i]], i.var = n, return.grid=TRUE, type="response")
      get.min.max <- filter(mm.df, wetland.bricknames == n)
      mn <- get.min.max[,2]
      mx <- get.min.max[,3]
      xout <- seq(mn, mx, (mx-mn)/1000)
      d <- approx(d[,1], d[,2], xout = xout, rule=2)
      d <- as.data.frame(d)
      response.df <- cbind(response.df,d)
   }

   #model stats
   AUC[i] <- fit[[i]]$cv.statistics$discrimination.mean
   dev[i] <- (fit[[i]]$self.statistics$mean.null - fit$self.statistics$mean.resid) / fit$self.statistics$mean.null

   print(paste0("done building model ", i))
}

importance <- rowMeans(df.importance)
imp.names <- c("ARI", "NDPOL", "NDWI", "REIP", "TPI", "TWI")
importance <- data.frame(imp.names, importance)
```

```
#------------------------------------------------------------
#------------------------------------------------------------
#####################################################################




#GENERATE RESPONSE CURVES AND MODEL STATS
#####################################################################
#------------------------------------------------------------
#------------------------------------------------------------
response.df <- response.df[,-1]
response.df.x <- response.df[,seq(1,length(response.df),2)]
response.df.x <- response.df.x[,1:length(wetland.bricknames)]
response.df.y <- response.df[,seq(2,length(response.df),2)]
for (i in 1:length(wetland.bricknames)){
   varname <- wetland.bricknames[i]
   collumns <- seq(i, length(response.df.y), length(wetland.bricknames))
   yvals <- response.df.y[,collumns]
   xvals <- response.df.x[,i]
   yvals.mean <- rowMeans(yvals)
   yvals.std <- apply(yvals,1,sd)
   yvals.neg.std <- yvals.mean - yvals.std
   yvals.add.std <- yvals.mean + yvals.std
   yvals.df <- cbind(xvals, yvals.mean, yvals.neg.std, yvals.add.std)
   yvals.df <- as.data.frame(yvals.df)
   if(i>0){
      xlim <- c(min(xvals), max(xvals))
   } else{
      xlim <- c(min(xvals), 1200)
   }
   g <- ggplot(yvals.df, aes(x=xvals, y=yvals.mean)) +
      theme_minimal()+
      geom_ribbon(aes(ymin=yvals.neg.std, ymax=yvals.add.std), fill="#6baed6", alpha=0.35) +
      geom_line(colour="#08519c", size=1.8) +
      xlab(varname) + ylab("predicted probability") + xlim(xlim) +
      theme(axis.title.x = element_text(size=22), axis.title.y = element_text(size=20),axis.text = element_text(size=16))
   assign(paste0(varname,".plot"), g)
}

#output model stats
setwd(outputs)
AUC <- mean(AUC)
dev <- mean(dev)
stats <- cbind(AUC,dev)
write.csv(stats, paste0("wetlandModelStats", ".csv"), row.names=FALSE)

tiff(paste0("wetlandResponseCurves", ".tiff"), width = 1200, height = 1000)
multiplot(ARI.plot, NDPOL.plot, NDWI.plot, REIP.plot, TPI.plot, TWI.plot, cols=2)
dev.off()

tiff(paste0("wetlandVarImportance", ".tiff"), width = 1000, height = 700)
ggplot(importance,aes(x=reorder(imp.names,-importance),y=importance)) + geom_bar(stat="identity", show.legend=FALSE, fill= "grey60") + theme_minimal() +
   theme(axis.title.x = element_text(size=24), axis.title.y = element_text(size=24),axis.text = element_text(size=22), legend.text = element_text(size=20), legend.title =
element_text(size=22)) +
   labs(x = "Variables") + labs(y = "Importance" )
dev.off()
#------------------------------------------------------------
#------------------------------------------------------------
#####################################################################



#####################################################################
#SECTION 2 PREDICT WETLAND PROBABILITY BY TILE
#####################################################################

#LOOP THROUGH PREDICTION OF RASTERS BASED ON MODEL FITS
#####################################################################
#------------------------------------------------------------
#------------------------------------------------------------

PUs <- readOGR("J:/LandCover/CurrentSurfaceWater/Boreal", "Boreal_PUs")
for (i in 1:length(PUs)){

   t1 <- Sys.time()

   #build raster brick
   setwd(location)
   fls <- wetland.tifs
   PU <- PUs[i,]
   #build raster brick
```

```
  r1 <- raster(fls[1])
  r1 <- crop(r1, PU)
  print("done cropping1")

  r2 <- raster(fls[2])
  r2 <- crop(r2, PU)

  r3 <- raster(fls[3])
  r3 <- crop(r3, PU)

  r4 <- raster(fls[4])
  r4 <- crop(r4, PU)

  r5 <- raster(fls[5])
  r5 <- crop(r5, PU)

  r6 <- raster(fls[6])
  r6 <- crop(r6, PU)


  r.b <- brick(r1,r2,r3,r4, r5, r6)
  names(r.b) <- wetland.bricknames

  #predict fit[[1]]
  beginCluster(20)
  r.b.p <- clusterR(r.b, raster::predict, args = list(model = fit[[1]], type = "response", n.trees = fit[[1]]$gbm.call$best.trees))
  endCluster()
  plot(r.b.p)

  #START PREDICTION OF RASTER STACK
  for (n in 2:length(fit)){
    beginCluster(20)
    prediction <- clusterR(r.b, raster::predict, args = list(model = fit[[n]], type = "response", n.trees = fit[[n]]$gbm.call$best.trees))
    endCluster()
    r.b.p <- stack(r.b.p, prediction)
    print(paste0("Done model prediction ", n))
  }

  wet <- calc(r.b.p, fun = mean)

  #Save wetland prediction rasters
  setwd(outputs)
  writeRaster(wet, paste0("WetlandProbability", i, ".tif"), datatype= "FLT4S")

  t2 <- Sys.time()
  t.diff <- difftime(t2,t1, units="hours")
  print(paste0("Done predicting PU ", i, " it took ", round(t.diff, 2), " hours"))

}
#-----------------------------------------------------
#-----------------------------------------------------
######################################################################
tt2 <- Sys.time()
t.diff <- difftime(tt2,tt1, units="hours")
print(paste0("total script took ", round(t.diff, 2), " hours"))
```

Script 1: R machine learning algorithm for classifying wetland occurrence in the boreal region.

2.5 Cross validation accuracy assessment

An independent cross validation accuracy assessment of the binary wetland/upland layers was completed by generating 200,000 points in 3x7 areas devoid of human footprint or surface water. Values from the 3x7 training data and the modeled wetland occurrence were then extracted for each point. With this data a traditional accuracy was calculated along with a confusion matrix, and a kappa statistic.

2.6 Additional processing

Once wetland probability was predicted across the boreal, areas with surface water (ABMI Boreal surface water inventory, 2017) were given a wetland probability value of 1 while areas with human footprint types seen in Table 4 (ABMI Human Footprint Inventory (HFI) 2014, 2017) were given a wetland probability value of 0. The final probability raster was converted into a binary wetland/upland grid. Another smoothed version of this binary grid was produced by using a 5x5 pixel majority filter.

Table 4: List of HFI 2014 sublayers used to assign a wetland probability of 0.

Human Footprint Inventory 2014 sublayer
Borrow Pits, Sumps, Dugouts, and Lagoons
Roads
Railways
Canals
Verge
Mine sites
Industrial sites
Well sites
Landfill
Other veg surfaces
Wind generation facilities
High density livestock operation
Residential areas
Cultivation

3 Results

The results of the BRT model show that TWI was the most important input variable (Figure 4). The two vegetation indices (ARI and REIP) proved to be the second and third most important variables. TPI, NDWI, and NDPOL had the least influence on the model but they all contributed over 7.5% to the model building. Figure 5 shows the mean partial dependence response curves for all six input variables and the standard deviation around the mean of the 40 models.

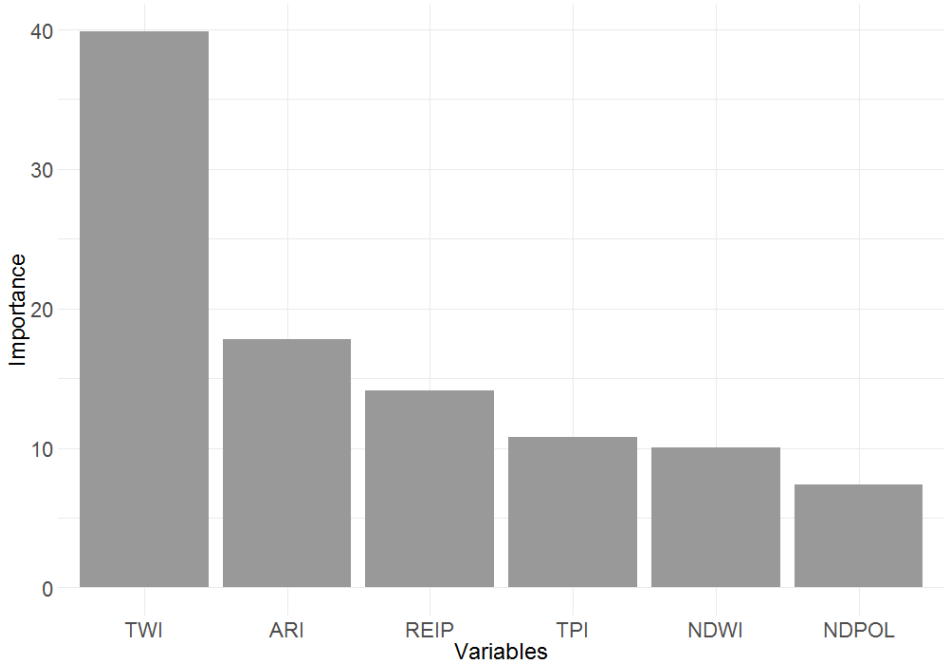


Figure 4: BRT model variable importance of all six input variables.

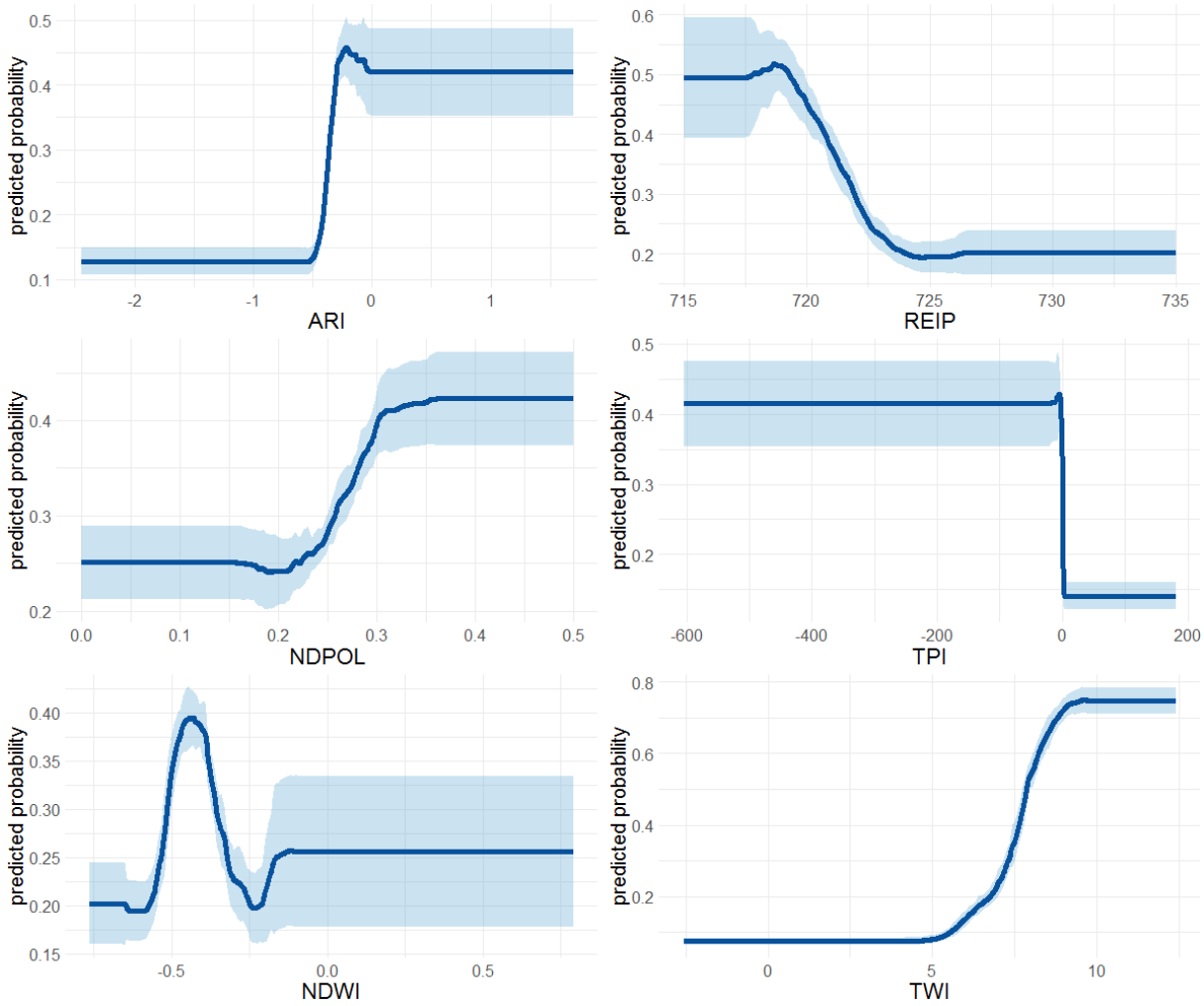


Figure 5: Mean partial dependence response curves of all six input variables for the wetland model. Predicted probability is the estimated probability of wetland occurrence at the given input value with all other variables held at their mean.

Figure 6 shows the model predicted across the boreal landscape. Blues represent high wetland probability and browns represent high upland probability. Low probabilities of wetland occurrence in the south are seen mainly due to heavy cultivation. Open water is taken from the ABMI boreal surface water inventory and given a value of 1 in this data set.

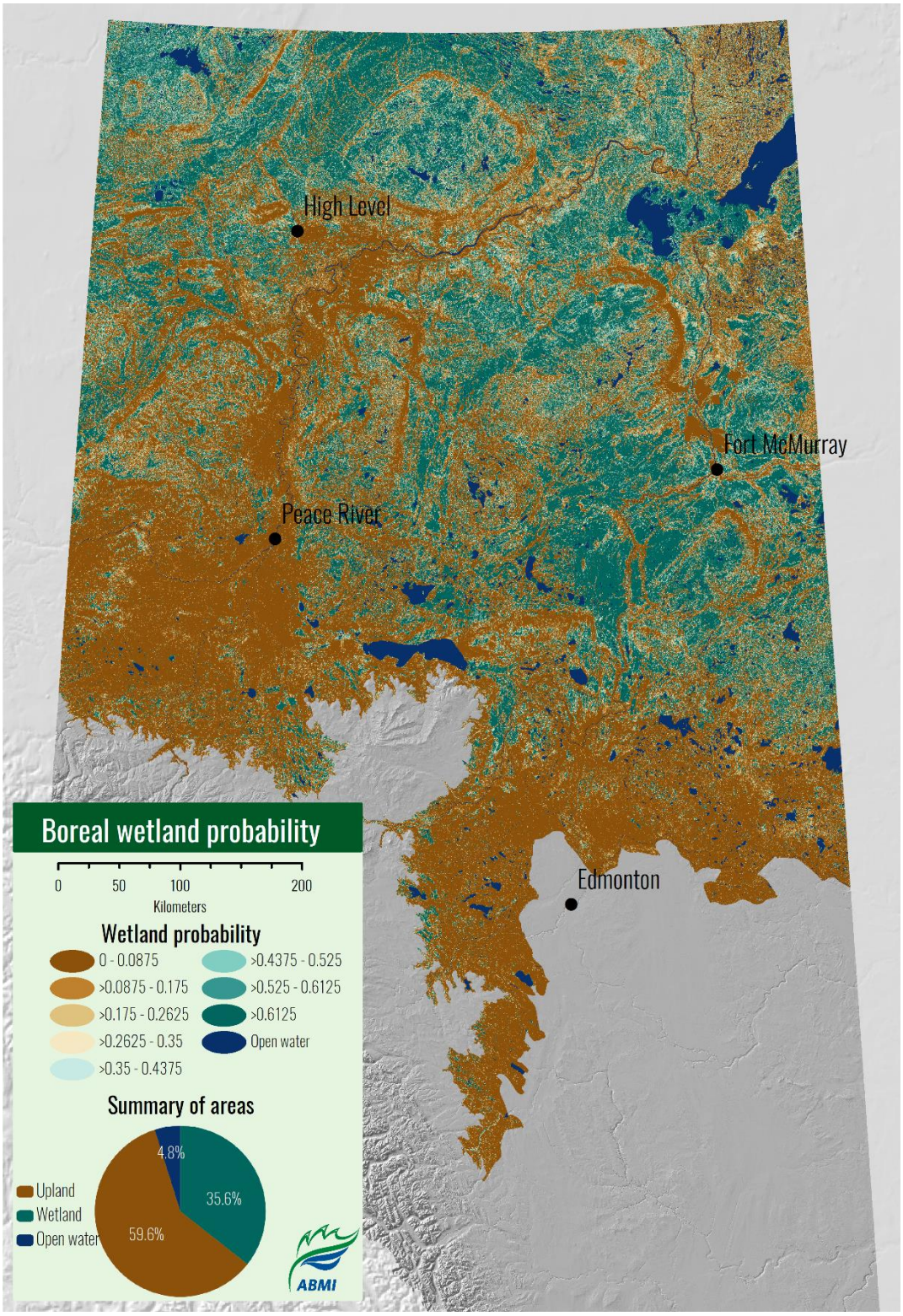


Figure 6: Wetland probability for the boreal region. Wetlands make up 35.6% of the study area while uplands and open water make up 59.6 and 4.8% respectively.

Figure 7 shows different error rates and kappa statistics for different probability thresholds. The lowest error rate (0.166) and highest kappa statistic (0.667) occur at a probability threshold of 0.35. Therefore, when completing a binary wetland upland classification the 0.35 probability was used to differentiate the classes. Table 5 and 6 show the results of the cross validation accuracy assessment using this 0.35 threshold as the distinction between wetland and upland. The overall accuracy to the 3x7s was seen to be 83.4%, the kappa statistic was 0.667, and the AUROC of the BRT model was 0.910.

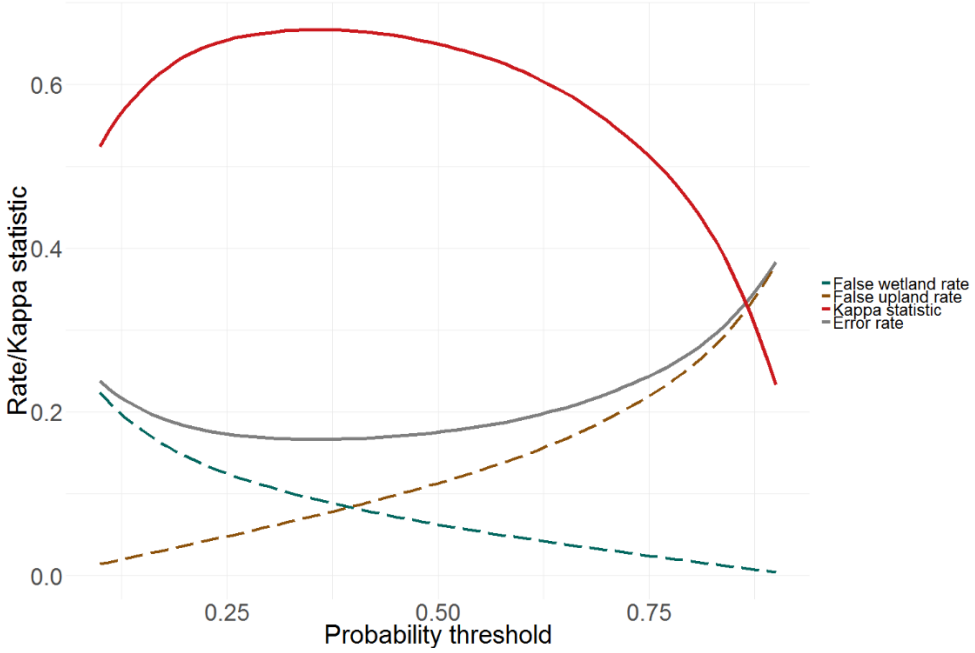


Figure 7: Error rate and kappa statistic at different probability thresholds. A wetland probability of 0.35 results in the lowest error rate and highest kappa statistic when classifying wetlands.

Table 5: Confusion matrix for the cross validation accuracy assessment to the3x7s. Overall accuracy shown in the bottom right in bold.

	Upland	Wetland	User Accuracy
Upland	81,140	14,433	0.849
Wetland	18,860	85,566	0.819
Producer accuracy	0.811	0.856	**0.834**

Table 6: Kappa statistic and AUROC of model and results

Kappa statistics of 3x7 cross validation	0.667
AUROC of boosted regression tree model	0.91

4 Conclusion

This document presents an automated framework for predicting wetland occurrence across large areas. The results show that the modelling framework can predict wetland occurrence across the entire boreal with an accuracy of 83.4%. Due to the automated nature of this framework and its use of cloud computing technologies, the process for predicting wetland occurrence takes about 4-6 weeks. Therefore it is now possible to produce consistent, repeated updates to wetland inventories for large areas of Alberta.

5 References

Alberta Biodiversity Monitoring Institute Remote Sensing Group. 2016. "ABMI Photo-Plot Quality Control Manual." Edmonton, Alberta.

Alberta Environment and Sustainable Resource Development. 2013. "Alberta Wetland Policy." Edmonton, Alberta, Canada.

Alberta Environment and Parks Alberta Merged Wetland Inventory https://geodiscover.alberta.ca/geoportal/catalog/search/resource/details.page?uuid=%7BA73F5AE1-4677-4731-B3F6-700743A96C97%7D

Alberta Environment and Sustainable Resource Development or Alberta Environemtn and Parks. 2012. "Primary Land and Vegetation Invontory (PLVI) - Standards and Specifications." Edmonton, Alberta.

Alberta Environment and Sustainable Resource Development or Alberta Environment and Parks. 2016. "Alberta Vegetation Inventory Extended." Government of Alberta. Edmonton, Alberta.

Böhner, J., Köthe, R., Conrad, O., Gross, J., Ringeler, A., and Selige, T. 2002. "Soil regionalisation by mean of terrain analysis and process parameterization." *European Soil Bureau*, Research Report No.7.

Brisco, B. 2015. "Mapping and Monitoring Surface Water and Wetlands with Synthetic Aperture Radar." In: *Remote Sensing of Wetlands: Applications and Advances*. Bocca Raton, FL, USA. Taylor and Francis Group.

Conrad, O., Bechtel, B. Bock, M., Dietrich, H., Fisher, E., Gerlitz, L., Wehberg, J., Wichmann, V., and Böhner, J. 2015. "System for Automated Geoscientific Analyses (SAGA) v. 2.1.4." *Geoscientific Model Development.* Vol. 8, pp. 1991-2007.

Copernicus Sentinel-1 data [2016, 2017], European Space Agency.

Difebo, A., Richardson, M., and Price, J. 2015. "Fusion of Multispectral Imagery and LiDAR Digital Terrain Derivatives for Ecosystem Mapping and Morphological Characterization of a Northern Peatland Complex". In: *Remote Sensing of Wetlands: Applications and Advances*. Bocca Raton, FL, USA. Taylor and Francis Group.

Ducks Unlimited. 2017. "Wetlands." www.ducks.ca.

Ducks Unlimited Canada. 2011. "A User's Guide to the Enhanced Wetland Classification for the ALPAC Boreal Conservation Project.

Elith, J., Leathwick, J.R., and Hastie, T. 2008. "A working guide to boosted regression trees." *Journal of Animal Ecology*, Vol. 77(No.4): pp. 802-813.

Gauthier, Y., Bernier, and Fortin, J-P. 1998. Aspect and incidence angle sensitivity in ERS-1 SAR data. *International Journal of Remote Sensing*, Vol. 19(No.10): pp. 2001-2006. *Animal Ecology*. Vol. 77(No.4), pp. 802-813.

Gitelson, A., Merzlyak, M., and Chivkunova, O. 2001. "Optical Properties and Nondestructive Estimation of Anthocyanin Content in Plant Leaves." *Photochemistry and Photobiology*, 71: 38-45.

Google Earth Engine Team. 2015. "Google Earth Engine: A planetary-scale geospatial analysis platform." https://eathengine.google.com.

Government of Alberta. 2006. Provincial LiDAR dataset. Edmonton, Alberta.

Hatfield, J.L., and Prueger, J.H. 2010. "Value of Using Different Vegetative Indices to Quantify Agricultural Crop Characteristics at Different Growth Stages." *Remote Sensing*, Vol. 2: pp. 562-578.

Herrman, I., Pimstien, A., Karnieli, A., Cohen, Y., Alchanatis, V., Bonfil, D.J. 2011. "LAI assessment of wheat and potato crops by VENμS and Sentinel-2 bands." *Remote Sensing of Environment*, Vol. 115 (No.8): pp. 2141-2151.

Hird, J., DeLancey, E.R., McDermid, G.J., and Kariyeva, J. 2017. "Google Earth Engine, Open-Access Satellite Data, and Machine Learning in Support of Large-Area Probabilistic Wetland Mapping." *Remote Sensing*, Vol. 9(No.12): pp. 1315.

Lee, J-S., Wen, J-H., Ainsworth, T.L., Chen, K-S., and Chen, A.J. 2008. Improved Sigma Filter for Speckle Filtering of SAR Imagery. *IEEE Transactions on Geosciences and Remote Sensing*, Vol. 47(No.1): pp.202-213.

McFeeters, S.K. 1996. "The use of the Normalized Difference Water Index (NDWI) in the delineation of open water features." *Remote Sensing Letters,* Vol. 17(No.7): pp. 1425-1432.

Parisien, M.A., Parks, S.A., Krawchuk, M.A., Flannigan, M.D., Bowman, L.M., Moritz, M.A. 2011. Scale-dependent controls on the area burned in the boreal forest of Canada, 1980-2005. *Ecological Application*, Vol. 21: pp. 789-805.

R Core Team. 2013. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org/.

Rouse, J.W., Haas, R.H., Schell, J.A., and Deering, D.W. 1973. "Monitoring vegetation systems in the Great Plains with ERTS." *In 3rd Symposium,* NASA SP-351 I, pp. 309-317.

Saha, S., Moorthi, S., Wu, X., Wang, J., Nagdiga, S., Tripp, P., Behringer, D., Chuang, H-Y., Iredell, M., Ek, M., Yang, R., Mendez, M.P., Dool, H.v.d., Zhang, Q., Wang, W., Chen, M., and Becker, E. 2014. "The NCEP Climate Forecast System Version 2." *Journal of Climate*, Vol. 27: pp. 2185-2208.

Touzi, R., Gosselin, G., Li, J., and Brook, R. 2011. "Peatland subsurface water flow monitoring using polarimetric L-band ALOS." *In: Proccedings on POLINSAR'11*, Frascatti, Italy.

Weis, A. 2001. "Topographic position and landform analysis." *Poster presentation, ESRI User Conference*, San Diego, California, USA.