# READ ME

# Overview

The purpose of this report is to present the process of extrapolating ground ice labels from the ground ice extent to the entire Fairbanks CWPP region. The data is in the “FNSB\_CWPP\_ground\_ice\_extended” csv file. The data set contains 13,296,100 observations, each a 30m-by-30m land parcel, and the following variables as columns:

* XCoord: x-coordinate
* YCoord: y-coordinate
* ground\_ice\_value: the original categories of ground ice, available only in the ground ice extent
* ground\_ice\_extended:
  + Categories: “Water”, “Low”, “Medium”, “High”
  + “Water” category is where either ABoVE value is “Water”, or where the ground ice value indicated “water\_body”
  + For ground ice extent, “Low”, “Medium”, “High” are derived from the observed ground ice categories, for the rest of Fairbanks CWPP, the categories are extrapolated (see Figure 4)
* Low\_prob: estimation of the likelihood that the ground ice category is “Low” (see Figure 3)
* Medium\_prob: estimation of the likelihood that the ground ice category is “Medium” (see Figure 3)
* High\_prob: estimation of the likelihood that the ground ice category is “High” (see Figure 3)

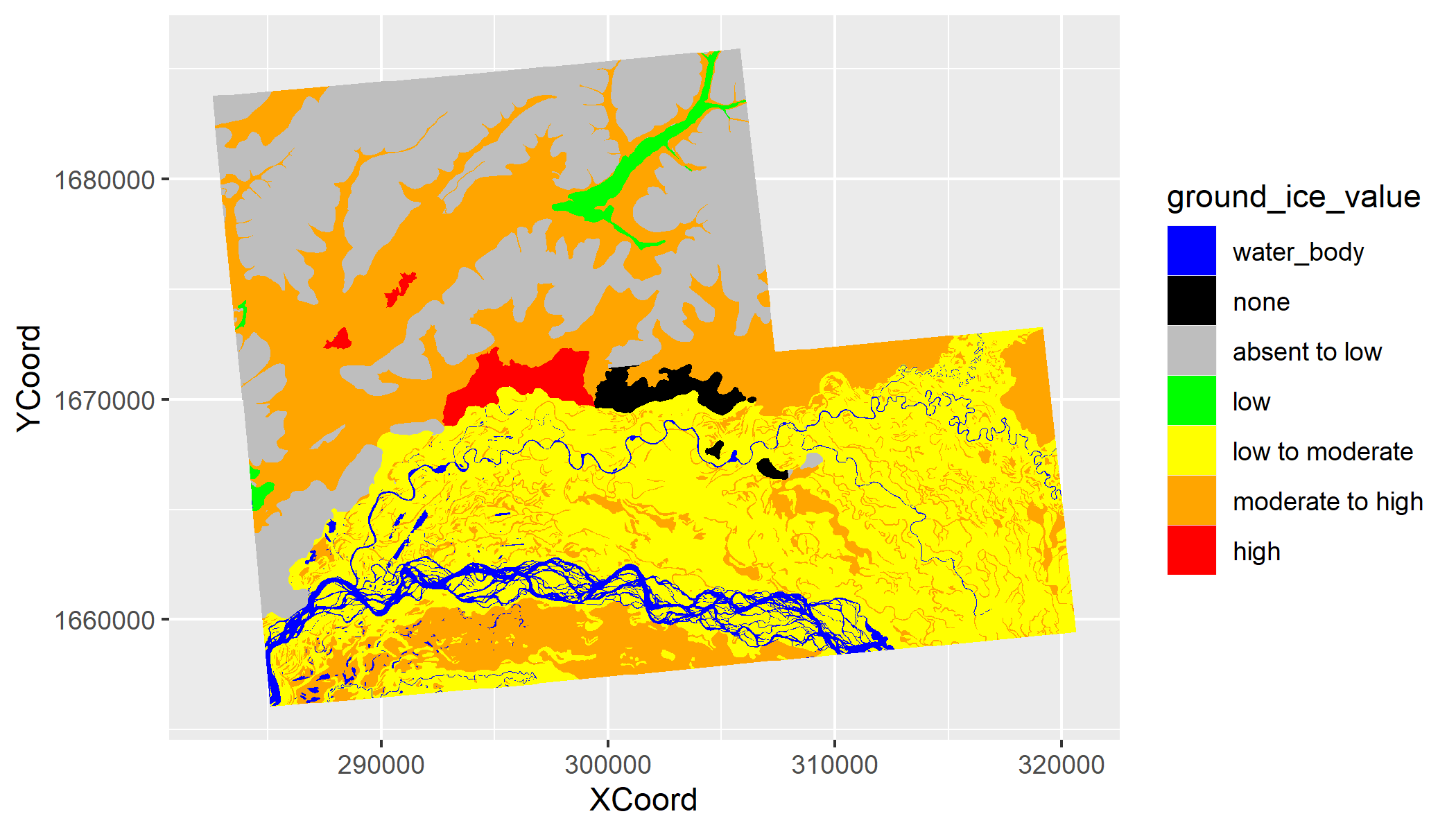
# Ground ice categories and source region for ground ice labels

We extrapolated a coarser scale of three categories of ground ice, which group the original Likert values of ground ice as follows:

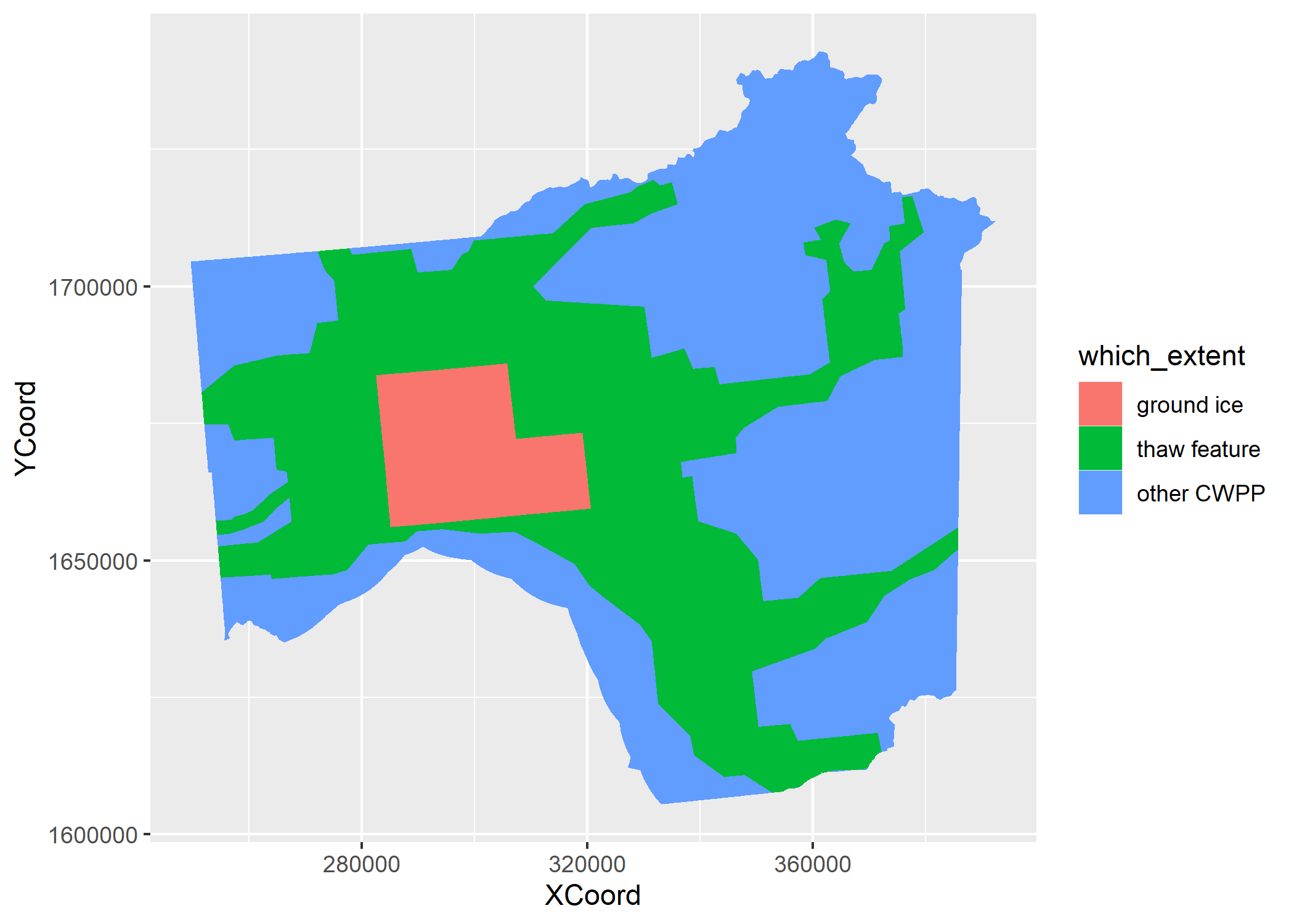
* Low = {"none, "absent to low", "low"} (29% of ground ice extent)
* Medium = {low to moderate"} (37% of ground ice extent)
* High = {"moderate to high", "high"} (34% of ground ice extent)

These groupings are reasonable based on looking at the spatial positions of the six Likert categories (Figure 1). Models trained on three balanced categories tend to be more reliable than models trained on six unbalanced categories.

We removed “water\_body” out of consideration, by relying on the ABoVE’s “Water” classification.



*Figure 1. Original ground ice values within the ground ice extent.*

Figure 2 shows in red the source region—ground ice extent—where the ground ice labels are known, in relation to the rest of the Fairbanks CWPP region (and the sub-region where thaw feature information remains available). 

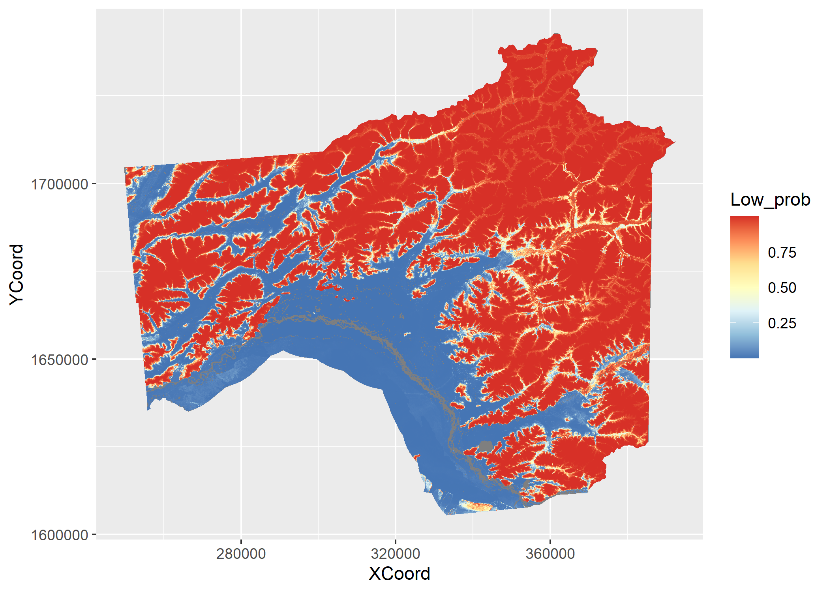
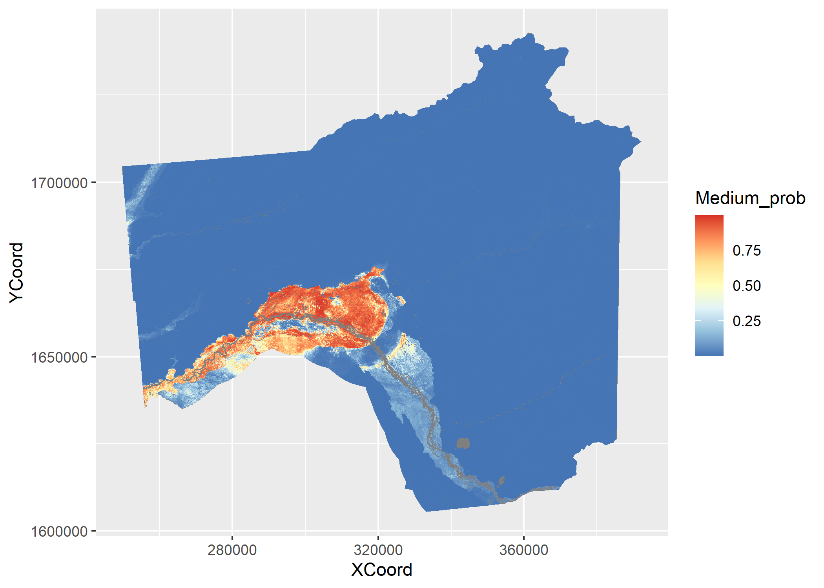
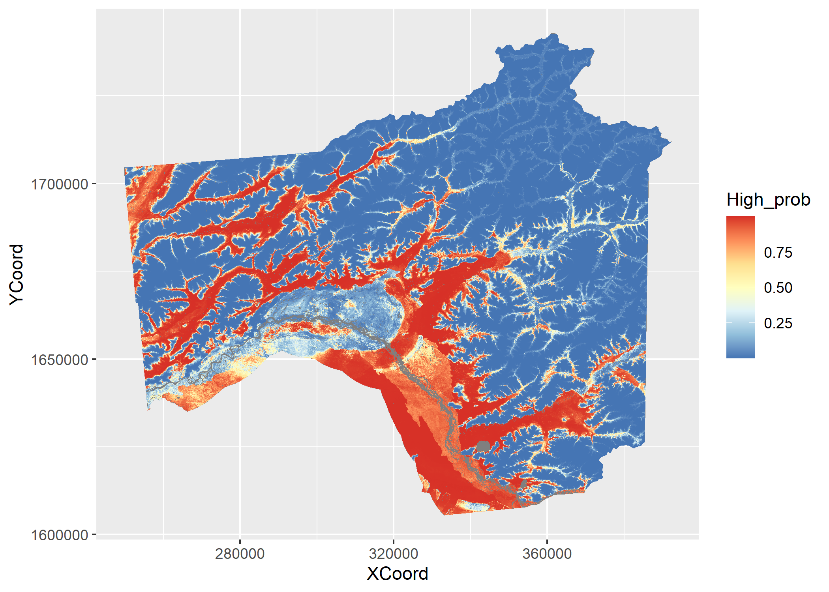
*Figure 2. Ground ice extent is the source region for the ground ice information; thaw feature extent is the region where thaw feature information is available, and thus can be used as a predictor. For the other CWPP region, the model cannot include thaw features as predictors.*

# Statistical learning model for extrapolating ground ice categories

WEhave trained two gradient-boosted decision tree aggregate models[[1]](#footnote-1)—one for thaw feature extent, and the other for outside the thaw feature extent—to predict which of the three ground ice labels a land parcel would have, based on the following features:

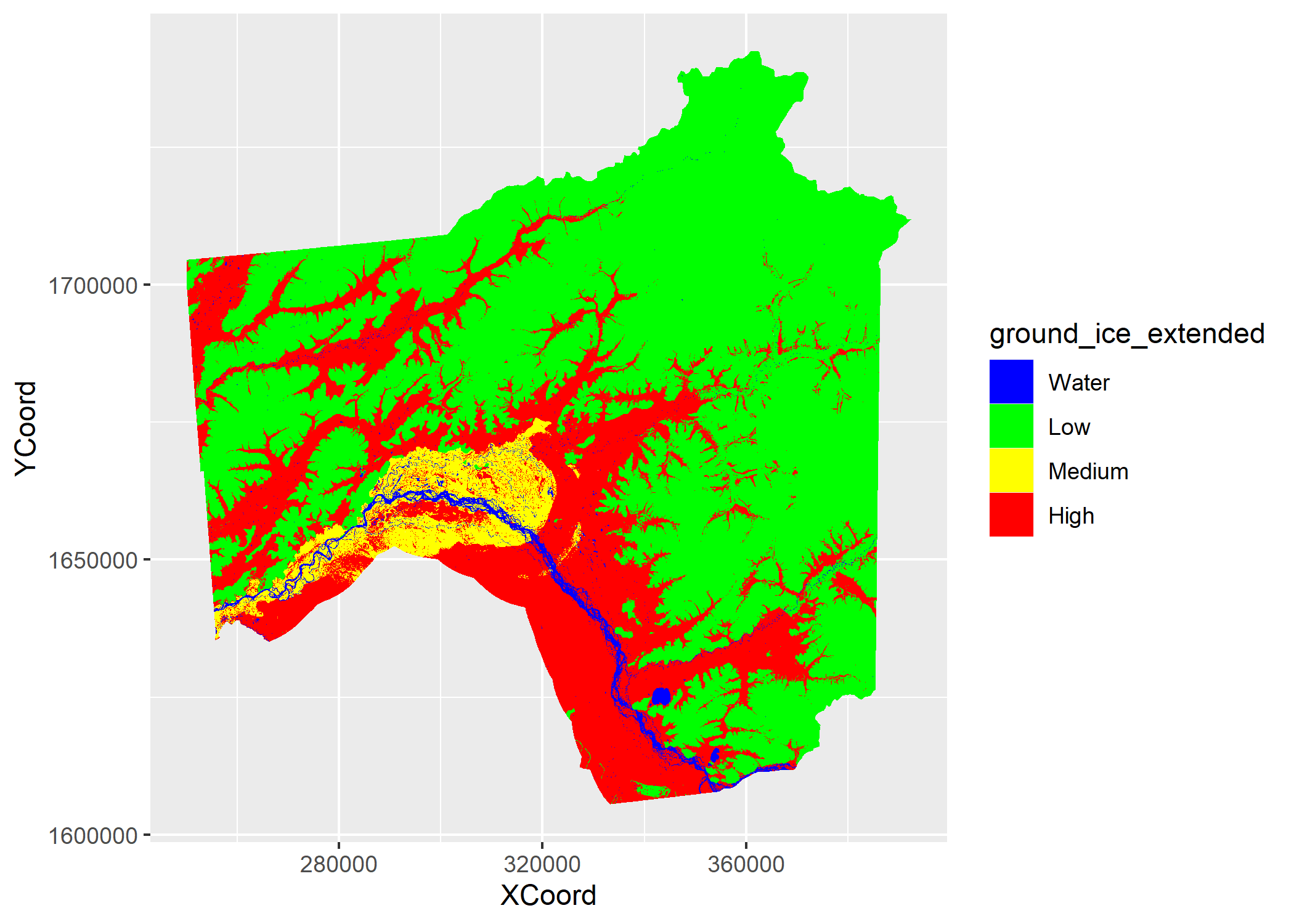
* Topography variables: elevation, local elevation (within 3km radius), slope
* Temperature: average summer and average winter temperatures for the decade of 1996-2005
* Tanana river floodplain indicator
* Wetland indicator (NWI-GEN)
* ABoVE vegetation categories from 1984 (adjusted to include “Black Spruce”)
* (for thaw feature extent only) ice wedge indicator

The output of such a model, given a new observation with all the predictor features, is a rough probability estimate[[2]](#footnote-2) of each of the three ground ice categories. Figure 3 shows these probabilities for each of the three ground ice categories. WEhave included estimates for the ground ice extent—the data on which both models got trained—to get a sense of accuracy.



*Figure 3. Probability estimates of the three ground ice categories—“High” top left, “Medium” top right, “Low” bottom—over the Fairbanks CWPP region.*

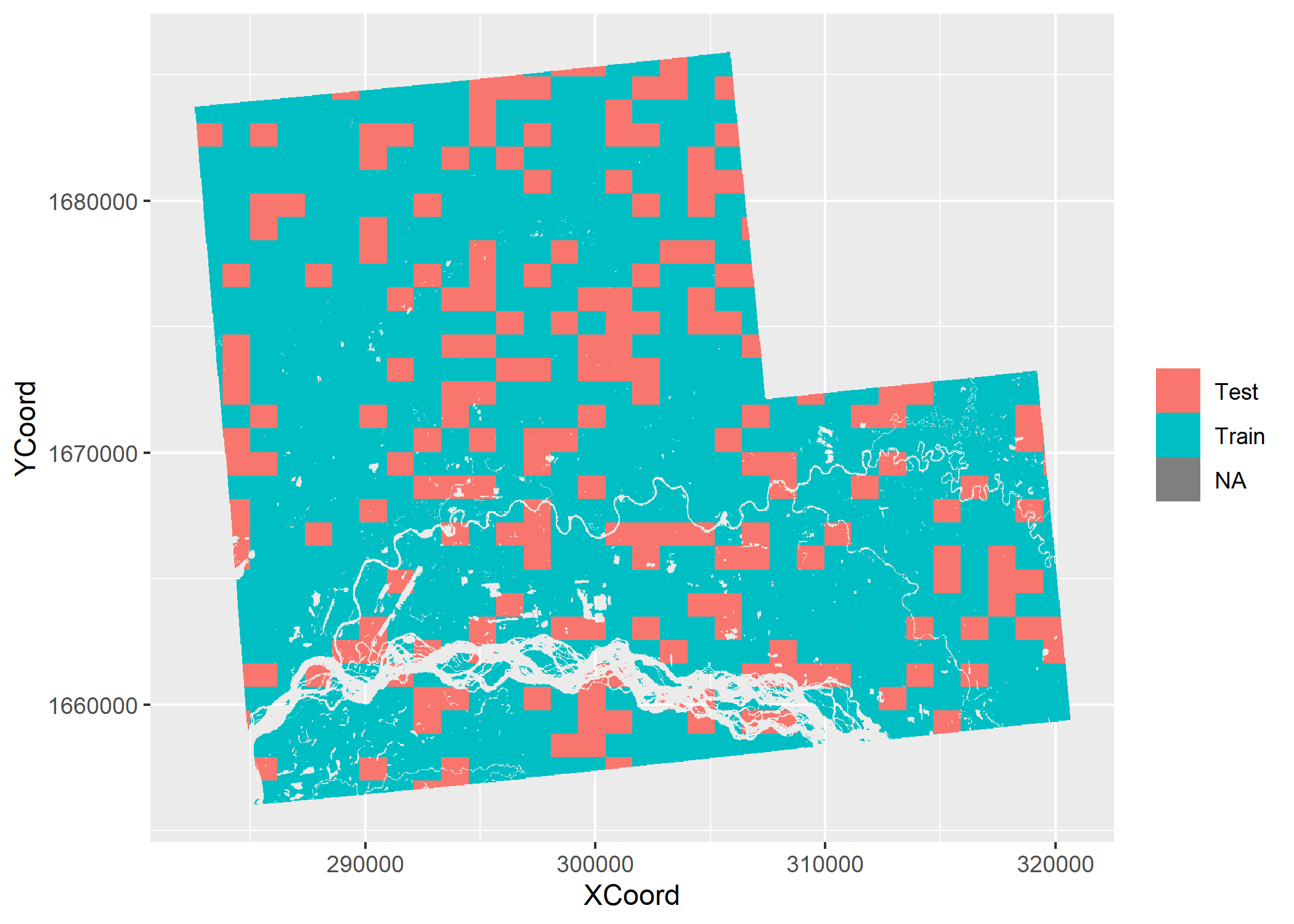
To decide between the three categories, we have chosen a cutoff for the “High” category that best balances precision and recall[[3]](#footnote-3) for “High”, otherwise choosing whichever category is highest. Figure 4 shows the resulting categories (the categories over the ground ice extent are the actual ground ice categories; everywhere else the categories are extrapolated). Dmitry and Louise should use their domain expertise to check if the extrapolated ground ice labels make sense.



*Figure 4. Ground ice categories extrapolated over the Fairbanks CWPP region.*

# Statistical learning model performance analysis

The most direct method for assessing how well a statistical model performs in predicting a categorical variable is to consider the “confusion” matrix on the test data—data that’s new to the model, but where the values for the categorical variable are known. My test data is comprised of randomly selected patches of land in the ground ice extent (about 25% of the total area, see Figure 5). Such a split between training and test data makes the testing more realistic. Nevertheless, since both the training and test data are bound within the peculiar ground ice extent region, the testing is still under idealized conditions where there is no substantial statistical difference between the testing ground and the training ground.



*Figure 5. De-correlated split of the ground ice region into training and test datasets. Observations with ABoVE’s “Water” classification are removed.*

The confusion matrix compares the model’s predictions with the actual categories on the data that’s new to the model. Table 1 shows the confusion matrix on test data for each of the models. Table 2 shows how good the models are at recognizing each category—for example, both models correctly identify 81% of the “High” categories. The high proportions on the diagonals indicate that both models tend to correctly recognize all three categories—though both are a bit worse at doing so for the “High” category, which they tend to confuse with the “Medium”. Table 3 shows how reliable the models are in their predictions—for example, both models are correct 87% of the time they predict the “High” category (so the false positive for “High” is 13%).

For the most part, WEwould say that the model works well on the test data (within the ground ice extent). There are other metrics that measure the performance of the models, and WEhave stuck a bunch of those in the Appendix, for the sake of record-keeping.

**Table 1.** Confusion matrices on test data for the thaw extent model (left), and for the model outside of thaw extent (right). The row specifies the category that the model predicted, the column specifies the actual ground ice category.

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | Thaw extent | | |  |  | Other CWPP | | |
|  | Actual categories | | |  |  | Actual categories | | |
| Prediction | Low | Medium | High |  | Prediction | Low | Medium | High |
| Low | 67,361 | 178 | 3,188 |  | Low | 67,012 | 193 | 3,299 |
| Medium | 120 | 64,916 | 10,277 |  | Medium | 124 | 64,919 | 10,244 |
| High | 4,529 | 3,871 | 58,116 |  | High | 4,874 | 3,853 | 58,038 |

**Table 2.**  Column percent of confusion matrices for the thaw extent model (left), and for the model outside of thaw extent (right). Diagonal cells correspond to correct predictions.

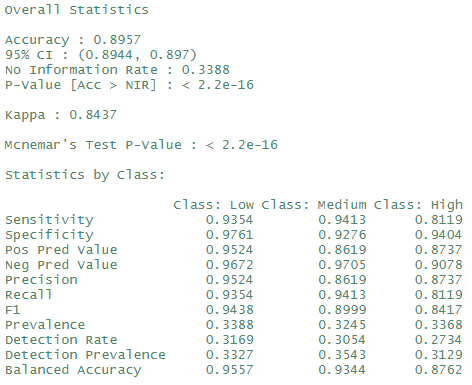


**Table 3.** Row percent of confusion matrices for the thaw extent model (left), and for the model outside of thaw extent (right). Diagonal cells correspond to correct predictions.

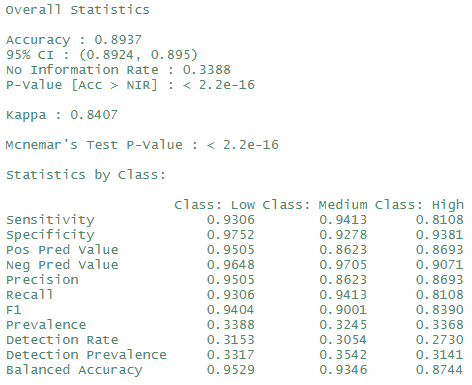


# Appendix

**Table 4.**  Performance statistics for thaw feature extent model on test data.



**Table 5.**  Performance statistics for other CWPP extent model on test data.



**Table 6.**  AUC-ROC measures.



*Figure 6. ROC curves for thaw feature extent model.*

|  |  |  |
| --- | --- | --- |
| Category | True positive v. false positive | Precision v. recall |
| Low |  |  |
| Medium |  |  |
| High |  |  |

*Figure 7. ROC curves for other CWPP extent model.*

|  |  |  |
| --- | --- | --- |
| Category | True positive v. false positive | Precision v. recall |
| Low |  |  |
| Medium |  |  |
| High |  |  |

1. We used the R package “xgboost”. Details are in the R script “2021-07-15 training xgboost script.R”. [↑](#footnote-ref-1)
2. Literally, how many of the aggregated decision trees voted for the category. [↑](#footnote-ref-2)
3. F1 measure. [↑](#footnote-ref-3)