**DOWNSCALING FOR THE 170 VALIDATION DAYS (SCRIPTS IN FOLDER)**

**Workflow**

Preliminary: need the following packages installed in R: fields, ranger, raster

Can try install.packages(“ranger”,”raster”,”fields”) in R, but this might not work. CU RC had to make a custom environment for Mitch with all installed, raster was particularly annoying. (Note from Mitch: need to install “Matrix” package before “ranger”)

The main workflow is below, numbered in order. Included is a summary about the output and where it is stored, along with changes that must be made when rerunning.

See twostep.sh for submitting a sequential run on Blanca, assuming all training sizes are identical.

1. twostep.growforest.R

* Grows the classification forest (in particular, a probability forest) and the regression forest based on training data
* Saves the forest objects in SierraBighorn/Rdata/forest/twostep/classification or SierraBighorn/Rdata/forest/regression. The final forests are classification/ranger.classifier.prob3e+05.Rda and regression/ranger.regression.prob3e05.Rda.
* Once this is run, the model is fitted and can now be used for downscaling (for this given training sample size)
* Changes: training size at the top of script

2. twostep.prediction.R [Mary Jo: This is the script to change and run to downscale for the 2000-2011 period]

* Loops through the 170 days and produces a pixelwise probability classification for each image as 0/btwn/100 fSCA, and then predicts the estimated fSCA values that are not 0% or 100%
* Probability tifs for btwn/100 are written out to SierraBighorn/downscaledv3/prob.btwn and SierraBighorn/downscaledv3/prob.hundred
* The regression predictions are performed on the full day, and saved in SierraBighorn/downscaledv3/regression
* These regression predictions are combined with the appropriate classifications, and the final two-stage .tif is written out to SierraBighorn/downscaledv3/downscaled
* Changes: training size at the top of script

3. SCA.comparedays.R

* This loops through the days and uses Karl’s SCAcompare.m script to produce daily summary statistics.
* Threshold currently is (0,0) and (0.15,0.15)
* The summary statistics are saved in SierraBighorn/Rdata/SCA as .csv files named by day
* Changes: training size at the top of script

**Notes**

All of the main R scripts follows the same format. The first part of each script is nearly identical and sets up the data in R: setting up proper dates for MODIS/landsat/satmask, reading in all predictor variables, loading forests, etc. After the setup comes the main computations: either forests are grown, or daily classifications/predictions/comparisons are made in a for loop. Tried to make most of the variable names self-explanatory.

Note that twostep.growforest.R is the only stage with training data, as the validation is performed on the whole dataset since the training data ultimately makes up such a small portion of the dataset. There are four tifs in total saved out: the twostep downscaled prediction, the regression downscaled prediction, the probability of (0,100)% fSCA, and the probability of 100% fSCA.

After these tif files are created, the script SCA.comparedays.R loops through the days, plots the estimated and true images, and uses Karl's SCAcompare function to compute summary statistics. Karl's original function was in Matlab, this R version is called SCAcompare.nobim.R and it doesn't save the large binary images but rather just the summary statistics.

For measuring the forest variables’ “importance”, would need to regrow with the option importance=TRUE. The ranger package gives error when predicting with a forest grown with this option, saying it is not recommended.

**Running on new 2000-2012 Data**

The important thing when extending to the new data is that the model will already be fit on our 170 validation days! So we only need worry about classification/prediction and not growing any forests.

The main change is to incorporate the new Landsat, saturation mask, and MODIS data in R. In particular, we would need to construct file different paths for the 2001-2012 landsat/satmask/MODIS file names. These new file paths would be used in the appropriate “for” loops in the main part of the classification/regression/comparison scripts, as we are now predicting on the new data rather than the 170 validation days.

The key part when making new file paths for landsat/satmask/MODIS is that the dates match. In the current script, most of the “header” portion is making sure that the dates of the following R variables are the same:

landsat.sc.file.names

modis.sc.file.names

sat.mask.file.names

These variables each contain 170 names of the paths to the data directories for the 170 validation days.

For the most part, the R scripts do not need to be changed and should be able to transition to the new dataset.

In summary, some file names would need to be changed to direct to the new data instead of the 170 validation days, and the for loops in twostep.prediction and SCA.comparedays would need to be modified to loop through this new data.

**Description of downscaledv3 and downscaledv3\_browse**

The downscaledv3\_browse is just a collection of .eps images that Will made. It looks like there are downscaled images as well as predicted class probabilities.

In downscaledv3, there are pictures I made in pix, you can ignore them, Will’s are visually better. The more interesting folders are downscaled, prob.btwn, prob.hundred, and regression. These folders are various downscaled predictions, and each have subdirectories 50000, 1e+05, 2e+05, 3e+05, which contain the .tifs at the 170 days but grown with trees from different daily sample sizes 50000, 1e+05, 2e+05, 3e+05 . The prob.btwn and prob.hundred are predicted downscaled probabilities of a pixel being classified as having some snow or 100% snow. And the regression is using the regression forest (not the classification forest) to predict fSCA. The downscaled folder is when we combine the classification forest and regression forest to get downscaled fSCA, which is probably what you are looking for.

So I think you are probably hunting all the .tif files in, say, downscaledv3/downscaled/3e+05