Wood River Water Collaborative Predictive Streamflow and Curtailment Date Model Details

Kendra Kaiser, n Boise State University, n Department of Geosciences

12/27/2021

Contents

1. Introduction

The Wood River Collaborative is a grassroots effort to tackle water usage challenges among irrigators, municipalities, and protect minimum flows for fish and wildlife habitat. Its many, basin-wide participants include private citizens, representatives of water agencies, non-profit organizations, private interests and the public sector. The outcome of the collaboration is to bring all stakeholders together and develop strategies and tools for best use of water for consumptive use, while conserving water for groundwater and in-stream flows.

The following suite of modeling tools were developed in response to stakeholder interests in improving management of surface and groundwater resources for agriculture and conservation purposes. These tools include automated data retrieval and organization for use in predictive models of irrigation season streamflow volume and timing in the Big Wood River Basin at the Hailey and Stanton Crossing gages, Camas Creek, and Silver Creek at Sportsman's Access (Figure 1). Curtailment dates for three priority water right dates are also predicted.

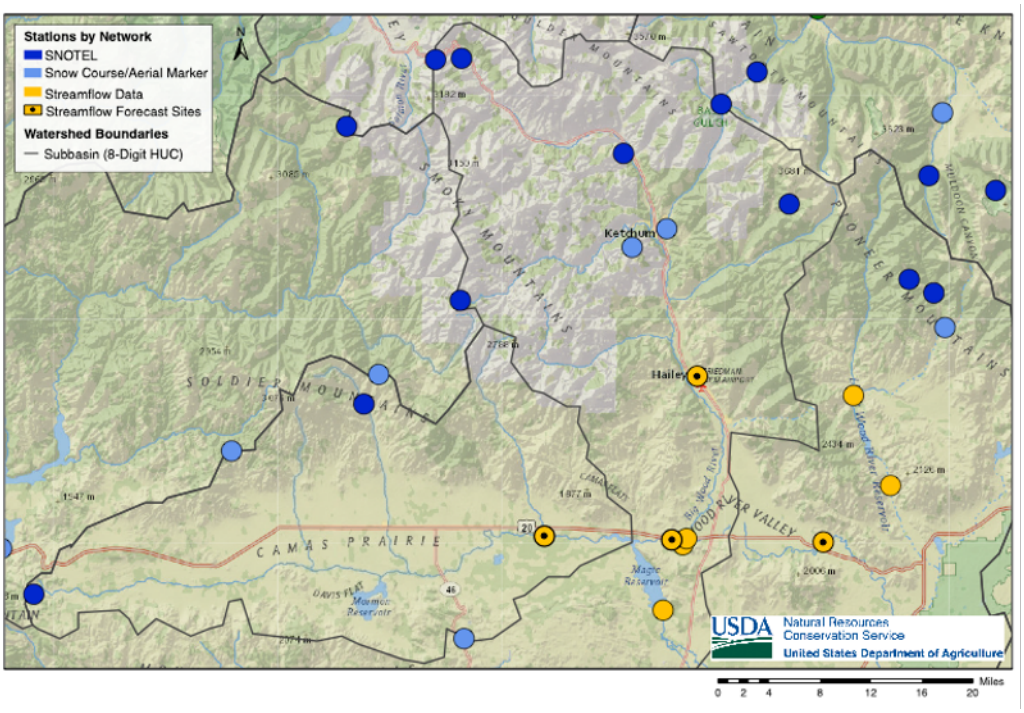


Figure 1: Map of the Big Wood River, Camas Creek and Silver Creek Watersheds and locations of automated data

Getting Started

Download code zip file by clicking on the green code button and save in a convenient place on your computer. You will need to enter this file path into the scripts to run the code locally.

Install RStudio on your computer https://rstudio.com/products/rstudio/download/

Setup file directories

There are two sets of file paths for the project. The first are the paths to where you saved the gitHub code files, and the second set is a local folder where final datasets and .csv files will be saved. The github file directories contain the code and a folder for output figures, these figures need to be in the same directory as the `.Rmd` file in order for the model run report to compile correctly. The local folder where output is saved

will be overwritten every time the model is run, so if saving model outputs is of interest, create another subfolder with a naming convention like '2021-02-01_ModelRun_lastname' where you copy and paste model outputs.

You will need to create the following set of subfolders in the local folder for data to be saved to: * April_output * data * February_output * March_output

```
#![alt text ><](https://github.com/kendrakaiser/WRWC/blob/master/figures/local_fil#e_dir.png?raw=true)
```

Run the Models

The only script you need to open to run the models is `run_models.R`. In this script you will modify the file paths and run date as described below. The `packages.R` script automatically installs and loads the libraries needed to run all the scripts. Once you have installed the packages once, you may comment out the installation lines with a `#`.

File paths

There are two sets of file paths for the project. The first are the paths to where you saved the gitHub code files, in the example: `'~/github/WRWC/`, and the second set is the local folder where final datasets and .csv files will be saved.

```
# GitHub File Paths
git_dir <<- '~/github/WRWC'

# Local File Paths
cd <<- '~/Desktop/WRWC'
```

Run Date and Prediction Year

The prediction year is the year of interest, and the 'run_date' is the date that you are running the models. There is a different set of models that are run for February, March and April. The optional name inputs are provided, any other naming convention will not work (e.g. March1 or mar1 will not work, only 'march1'. While the models were developed using data from th first of the month, they models can be run any date in the month.

```
# set prediction year
pred.yr <<- 2019
# set run date for set of models to use 'feb1', 'march1', 'april1'
run_date <<- 'april1'
```

Information for Model Run Report

This information is printed at the top of the model run report and is valuable for tracking model outputs over time.

```
# info for model run report
author = "Kendra Kaiser"
todays_date = "01/14/2021"
```

Model Support

If you run the model and an error occurs, the process for getting help is to raise an 'issue', you can do this by following these steps https://docs.github.com/en/free-pro-team@latest/github/managing-your-work-on-github/creating-an-issue. This will automatically send me an email so that I can help resolve the issue.

2. Methodology & Model Fits

Overview

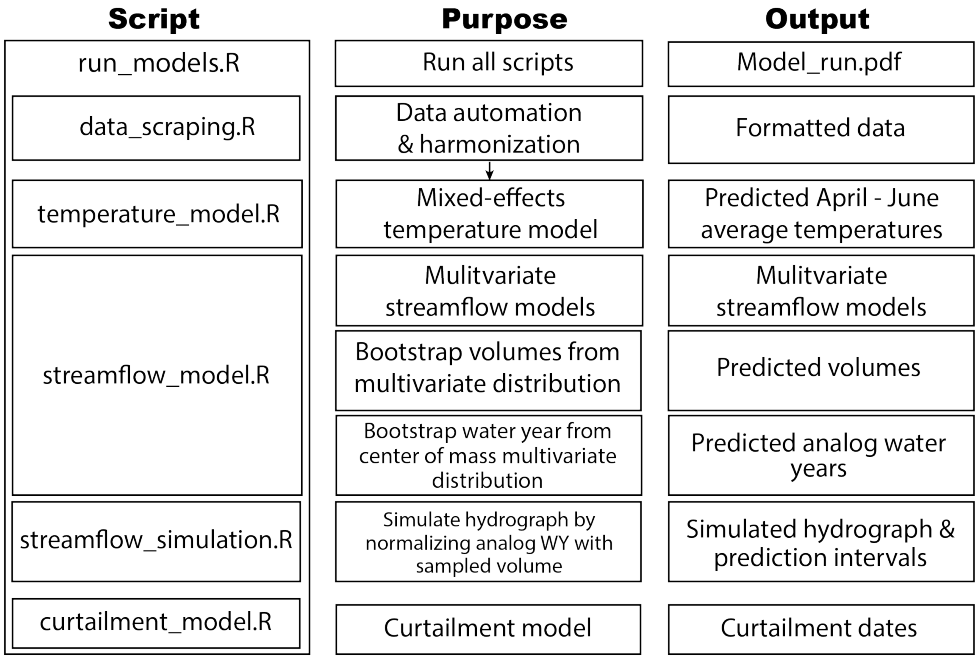
Individual multivariate linear regression models were developed for each of these locations using USGS streamflow data, Snotel SWE and temperature data, AgriMet temperature data (Table 1). The Baysian Information Criterion (BIC, 2008) was used to select model parameters for each of the gage locations for total irrigation season streamflow volume and timing as well as the selected curtailment dates. Timing is characterized by the center of mass which is the mean of the probability distribution of April - September streamflow, or the date of the "mean" streamflow between April and September.

Data	Source
Snow water equivalent (SWE)	NRCS SnoTel (9)
Streamflow	USGS gauges (4)
Temperature	NRCS AgriMet
Water rights & historic curtailment dates	District 37 Watermaster
Irrigation Diversions	District 37 Watermaster

Once the linear regression models were developed for total irrigation season volume and timing at each location, multivariate distributions were used to stochastically model hydrographs for each location. The residuals (standard error) from each of the regression models and correlations between gauge stations are used to create the multivariate distributions. This ensures that given a set of predictor variables (e.g. SWE, temp), the predicted volumes will be statistically consistent across gage locations (e.g. the models wont predict that Camas Creek will have really low runoff year while the Big Wood has a really high runoff year because they are statistically correlated). Repeated, random selection from these multivariate distributions produces a `sample` of predicted volumes and timing of streamflow. The samples of total volume and streamflow timing are then used to create simulations of the irrigation season hydrograph. Variability in final model outputs is quantified by percentiles of the resulting predictions.

The methods for predicting curtailments dates currently follow those for streamflow timing, where once the linear regressions are made the covariance between curtailment dates are used to create a multivariate distribution from which potential dates are sampled from.

The suite of linear regression models are unique to each run date, February 1st, March 1st and April 1st. In February and March, the linear regressions for diversions above the Big Wood at Stanton gage did not preform well, so the diversions are sampled from a normal distribution created from the historic data. The curtailment models are currently only in the April model run, this is largely due to the high uncertainty in the results of these models. This uncertainty largely comes from the compounding uncertainty from predictions of total seasonal streamflow volume, temperatures and diversions.



Reproducibility

All model scripts have been developed using gitHub as the code repository. This enables tracking of all model changes, sharing of model code with WRWC members and a mechanism for users to post 'issues' to the code repository (`https://github.com/kendrakaiser/WRWC`). When the model is updated a versioning standard will be used to update

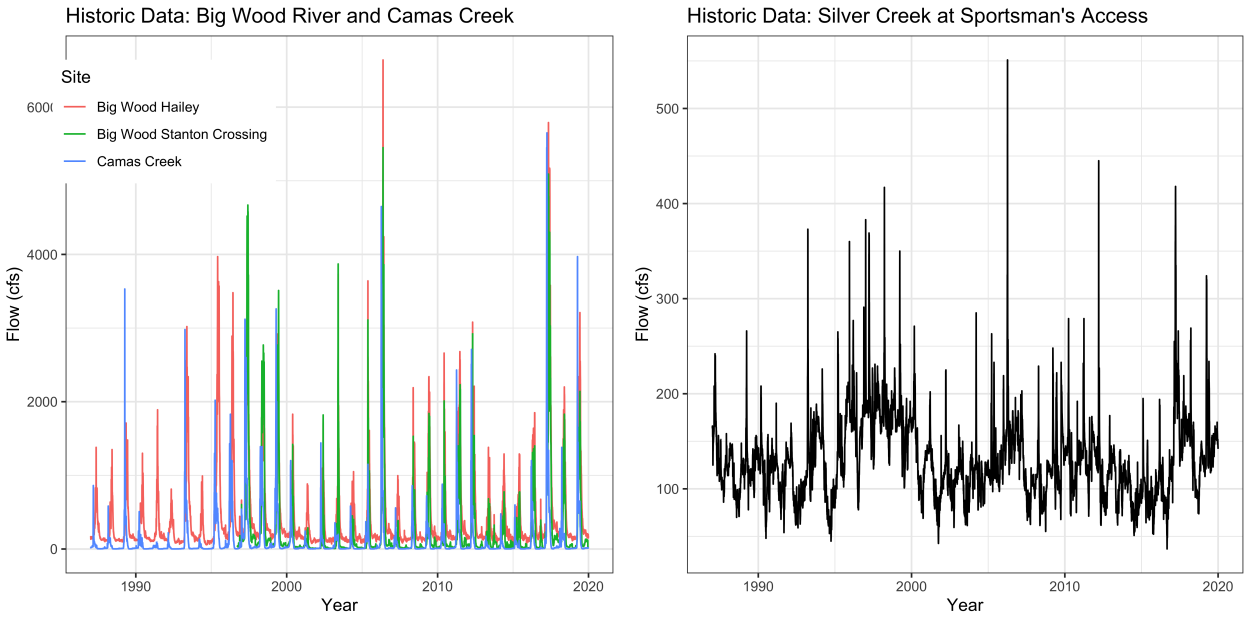
2.1 Data Downloading and Organization

Automation of data downloads and processing ensures that all data is formatted properly. Creating a local folder for each model run where all formatted data is saved will be valuable for reproducibility purposes.

Table 1: USGS Sites

	station_nm	huc_cd	begin_date	end_date	abv
1	BIG WOOD RIVER AT HAILEY ID TOTAL FLOW	17040219	2006-10-01	2021-12-12	bwb
2	BIG WOOD RIVER AT STANTON CROSSING NR BELLEVUE ID	17040219	1996-09-18	2021-12-12	bws
3	CAMAS CREEK NR BLAINE ID	17040220	1987-08-17	2021-12-12	cc
5	SILVER CREEK AT SPORTSMAN ACCESS NR PICABO ID	17040221	1987-08-18	2021-12-12	sc

USGS



Snotel

Snotel data from all locations in the Big Wood, Camas Creek and Little Wood drainages are included in the automated data downloading. This data includes snow water equivalent (SWE), cumulative precipitation, max, min and average daily temperatures.

```
kable(snotel_sites %>% dplyr::select(start, end, site_name, huc8, abv), caption = "Snotel Sites")
```

Agrimet

A specific function has been developed to download the AgriMet data without timing-out the servers. This has been added to the `code` folder (`grabAgriMetData.R`) to make installation easier. Temperature data from Fairfield and Picabo are included.

Snow Cover Extent

Remotely sensed snow cover extent was explored as a means to represent snow derived water availability in conjunction with Snotel data for the predictive streamflow model. Google Earth Engine (GEE) was used to extract snow covered extent (SCE) from Landsat images (16 day return interval, 30m resolution). For the purposes of this exploratory analysis, data from Landsat5 TM from 1983-2013 was used. A GEE script gathers all images over the Wood River Basin (WRB), filters out pixels that are cloud covered, or otherwise problematic, and applies the Normalized Difference Snow Index (NDSI) to the remaining pixels. Although this analysis lead to 408 total images that have greater than 50% coverage (clear pixels), very few of these images cover winter months. Additional modeling will be needed to use any remotely sensed snow cover data.

Table 2: Snotel Sites

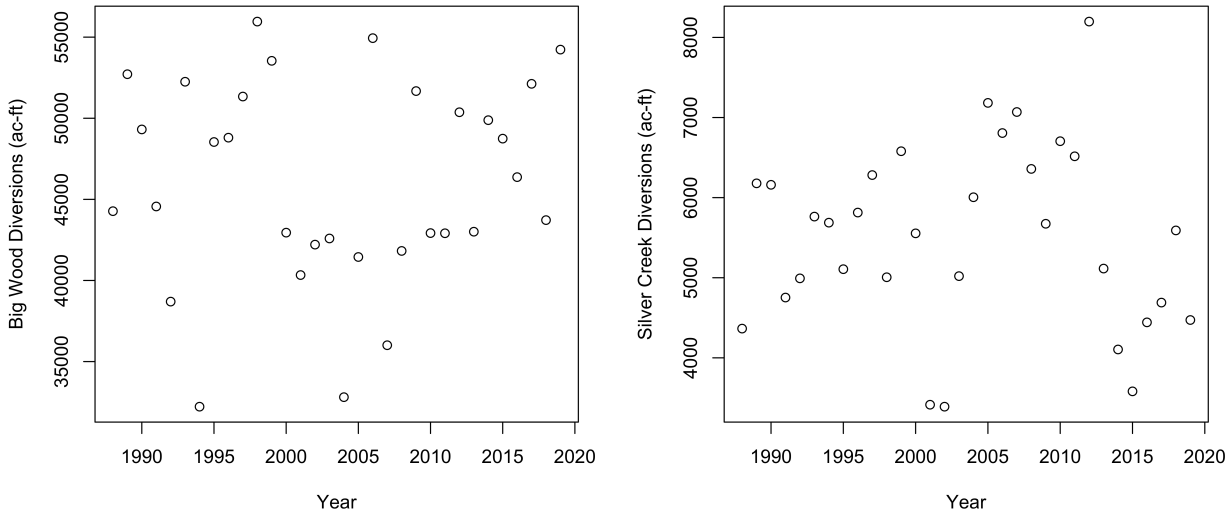
start	end	site_name	huc8	abv
1992-10-01	2021-12-12	chocolate gulch	219	cg.swe
1979-10-01	2021-12-12	galena	219	g.swe
1978-10-01	2021-12-12	galena summit	219	gs.swe
1979-10-01	2021-12-12	hyndman	219	hc.swe
1979-10-01	2021-12-12	lost-wood divide	219	lwd.swe
1979-10-01	2021-12-12	dollarhide summit	219	ds.swe
1991-10-01	2021-12-12	camas creek divide	220	ccd.swe
1985-10-01	2021-12-12	soldier r.s.	220	sr.swe
1979-10-01	2021-12-12	garfield r.s.	221	ga.swe
1978-10-01	2021-12-12	swede peak	221	sp.swe
1979-10-01	2021-12-12	stickney mill	303	sm.swe
1979-10-01	2021-12-12	bear canyon	101	bc.swe

Diversion & Curtailment Data

This data was compiled by WRWC by manually entering data from the irrigation district black books and should be updated annually for future model revisions. Currently the following diversions are included:

BWB: Tom P2, Lewis 1, Ketchum 2, McCoy 3, Peters 17C1, Hiawatha 22, Osborn24, and Cove 33 (above Hailey), WRVID 45, Bannon 49, Glendale 50, Baseline 55, Brown 57F1, Brown 57F2, Black 61, Graf 62, Uhrig 63, Flood 64 SC: Teeter Canyon P5, Stalker Creek P7, Gillihan Bashaw, Gillihan Picabo Live, Gillihan Woods, Stanfield 13, Albrethson 17, Kilpatrick 18, Iden 19 Fish and Game, Iden 19 Picabo Livestock

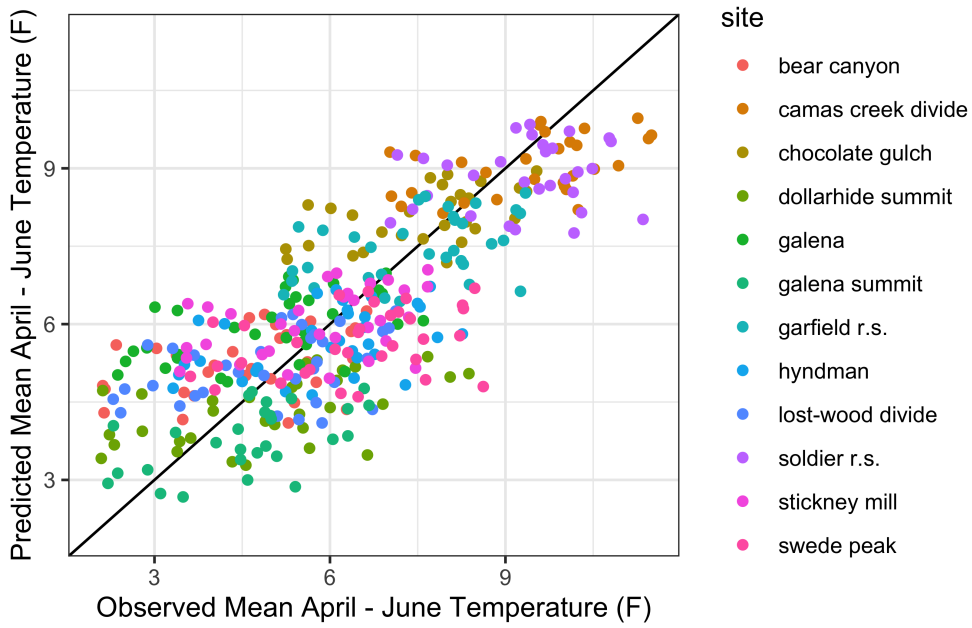
Smaller diversions were not included in this model version as they were considered to be minor given time constraints.



2.2 Temperature Model

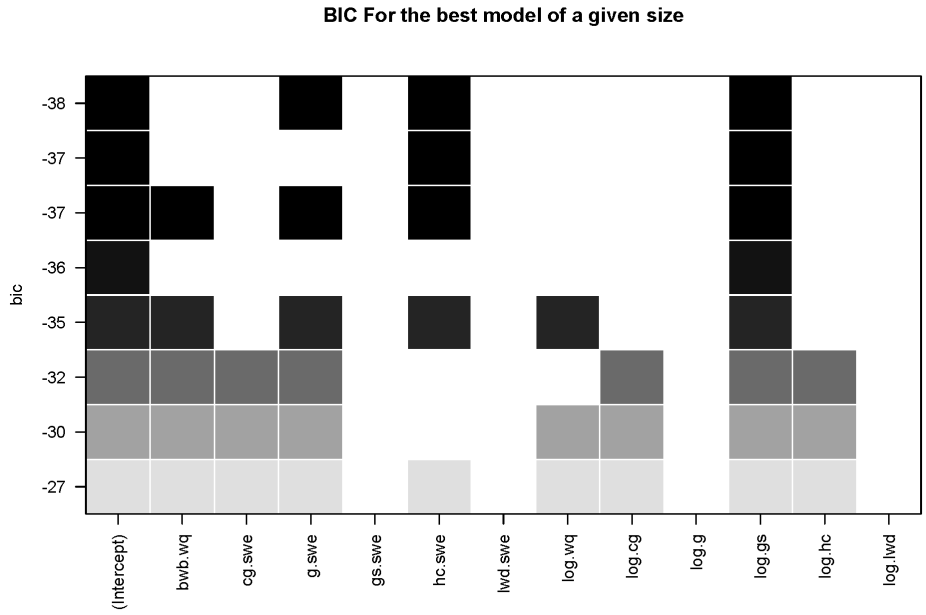
A linear regression model was developed to predict mean April - June temperatures at each SNOTEL site. April - June temperatures are predictors in the center of mass regressions, so a bootstrapped sample of predicted temperatures are used.

```
# linear regression using elevation and year
lr.elev<- lm(spring.tempF~  elev+year, data=input)
# covariance matrix (var-cov matrix)
site.cov<- cov(spring.tdata[-1], use="complete.obs")

#predict the mean april-june temperature at each site
new.data$spr.tempF[1:12]<-predict(lr.elev, new.data[1:12,])
# use the mean of fairfield and picabo - has no trend and decreasess strength of lm
new.data$spr.tempF[13]<- mean(tdata$spring.tempF[tdata$site == "fairfield"])
new.data$spr.tempF[14]<- mean(tdata$spring.tempF[tdata$site == "picabo"])

# Draw stream temperatures using multivariate normal distribution
nboot<-5000
aj.pred.temps<- mvrnorm(nboot, new.data$spr.tempF, site.cov)
write.csv(aj.pred.temps, file.path(data_dir, 'aj_pred.temps.csv'))
```



2.3.1 Streamflow Models

Models are automatically developed in the `streamflow_modes.R` script. The full suite of predictor variables are subset for each gage and the final set of predictor variables were determined using the `regsubsets` package, which enables visualization of adjusted R2 and BIC of each parameter set, and automatically saves the best predictor variables `vars<-reg_sum$which[which.min(reg_sum$bic),]`. Final streamflow models and center of mass models are saved in lists, such that the model structure and summary may be uploaded in other scripts. This workflow enables the models to be automatically updated every year.



The `streamflow_predictions.R` script creates predictions for each model for the user defined year. Data is imported, and data structures are set up to save model output. The `modOut` function returns relevant metrics and statistics from the modeled results for the year being predicted.

```
modOut<- function(mod, pred.dat, wq, vol, meanSWE, lastQ){
  '
  mod:      input model
  pred.dat: data.frame of prediction variables
  wq.cur:   this years winter baseflow
  wq:       array of historic winter flows (e.g. hist$cc.wq)
  vol:      array of historic april-sept volumes  (hist$cc.vol)
  hist.swe: mean(arrays of historic SWE from model snotel sites)
  lastQ:    last years summer streamflow volume (ac-ft)
  '
}
```

After this function is defined, the same set of steps occurs for each linear model. 1) The model parameters are subset from the full data set, 2) The prediction data is subset 3) Predictions are made, and outputs (estimated volume and standard error) are saved.

```
# 1. Subset Big Wood Variables
hist <- var[var$year < pred.yr,] %>% dplyr::select(c(bwb.vol, vol.params$bwh$vars)) %>% filter(complete
swe_cols <- hist %>% dplyr::select(contains('swe'))

#2. Subset Prediction Data
pred.dat<-var[var$year == pred.yr,] %>% dplyr::select(vol.params$bwh$vars)

#3. Make Big Wood at Hailey Predictions and Save Output
mod_sum[1,1]<-summary(vol.mods$bwh_mod)$adj.r.squared
mod_out<- modOut(vol.mods$bwh_mod, pred.dat, var$bwb.wq[var$year == pred.yr],
var$bwb.wq[var$year < pred.yr], hist$bwb.vol, mean(colMeans(swe_cols, na.rm=T)),
var$bwb.vol[var$year == pred.yr-1])
```

```
output.vol[1,] <- mod_out[[1]]
pred.params.vol[1,] <- mod_out[[2]]
```

After the streamflow volume model section of code, the same procedure is done for creating multivariate linear regressions for predicting center of mass. All model details and fits are provided in the `ModelFits.pdf`.

2.3.2 Streamflow Correlations

Given the proximity of the three basins, correlations between the basins' total annual irrigation season streamflow and center of mass allow us to ensure that the predicted flows at each gage are representative of how regional climatic patterns will be effecting all locations. For example, we would not expect Camas Creek to have an exceptionally dry year in a year when the Big Wood is experiencing an exceptionally high streamflow year. The correlation between sites is combined with the standard error from each linear model to create a covariance matrix which is use to bootstrap model predictions (or run the models multiple times).

```
# Correlation matrix between streamflow volumes, diversions and centers of mass
cor.mat<-cor(cbind(flow.data[c(1,3,5,7,9,10)],flow.data[c(2,4,6,8)]),use="pairwise.complete")
# Create covariance matrix by multiplying by each models standard error
# pred.pars[1,]: fitted values; pred.pars[2,]: sigma (standard error)
pred.pars<-rbind(pred.params.vol, pred.params.div, pred.params.cm)
outer.prod<-as.matrix(pred.pars[,2])%*%t(as.matrix(pred.pars[,2]))
cov.mat<-cor.mat*outer.prod
```

	bwb.vol	bws.vol	cc.vol	sc.vol	bwb.cm	bws.cm	cc.cm	sc.cm
bwb.vol	1	0.98	0.91	0.87	0.28	0.37	0.09	-0.32
bws.vol	0.98	1	0.92	0.9	0.28	0.37	0.11	-0.3
cc.vol	0.91	0.92	1	0.92	0.35	0.41	0.05	-0.45
sc.vol	0.87	0.9	0.92	1	0.4	0.45	0.16	-0.29
bwb.cm	0.28	0.28	0.35	0.4	1	0.97	0.65	-0.16
bws.cm	0.37	0.37	0.41	0.45	0.97	1	0.68	-0.18
cc.cm	0.09	0.11	0.05	0.16	0.65	0.68	1	0.13
sc.cm	-0.32	-0.3	-0.45	-0.29	-0.16	-0.18	0.13	1

Figure 2: Correlation matrix between gages

Flow volumes are then sampled from the multivariate distribution.

```
vol.pars<-rbind(pred.params.vol, pred.params.div) # only use predictions from volume models
vol.sample<-mvrnorm(n=5000,mu=(vol.pars[,1]),Sigma=cov.mat[1:5,1:5]) # historical covariance of volumes
```

This results in a distribution of potential volumes for each gage, given the input predictor variables.

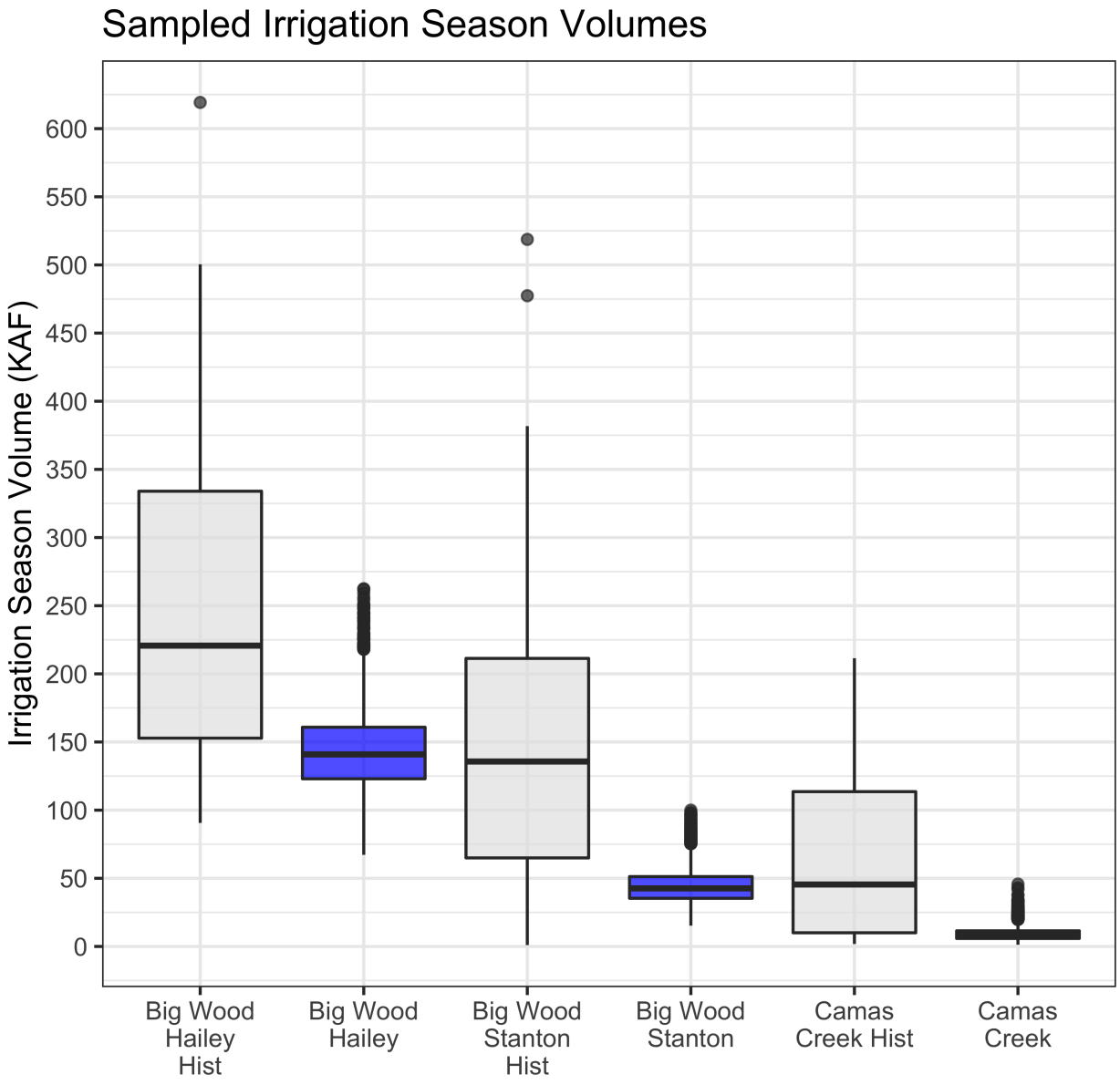


Figure 3: Distrubution of sampled volumes at each gage

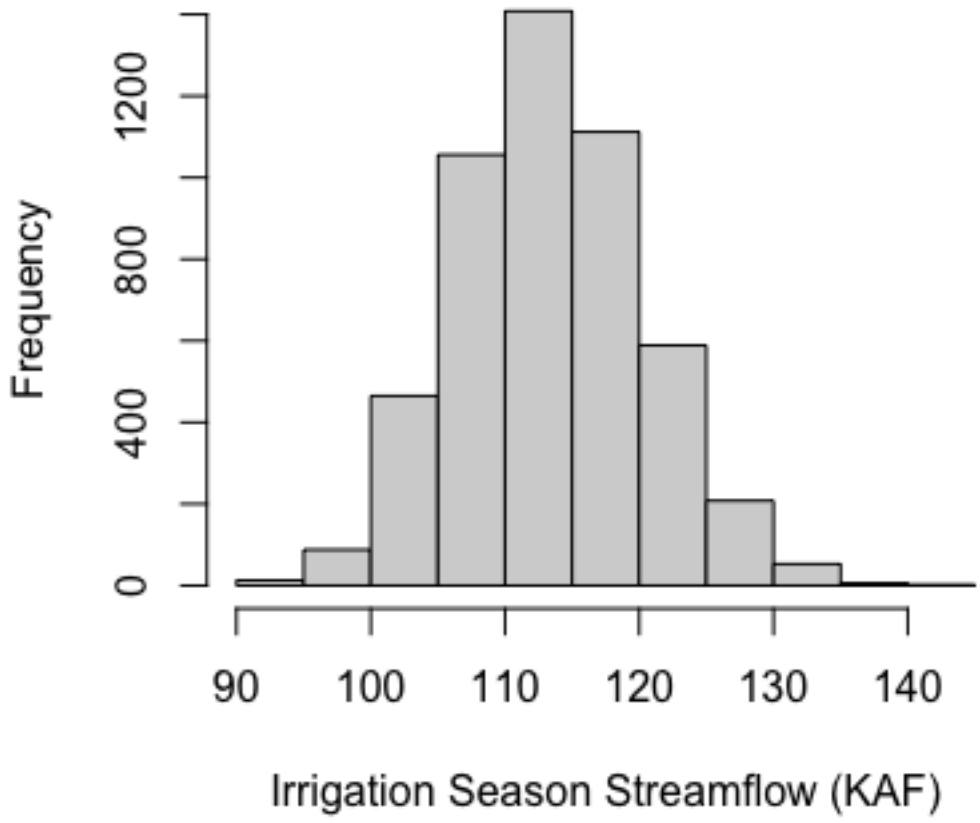


Figure 4: Distrubution of sampled volumes at each gage

A similar process is used for estimating the timing of runoff.

```
cm.data = var[var$year >= 1997 & var$year < pred.yr,] # only use complete dataset
cm.data = cm.data %>% select(year, bwb.cm.nat, bws.cm.nat,cc.cm, sc.cm)
cm.data$prob<-NA

# pmvnorm calculates the distribution function of the multivariate normal distribution
for(i in 1:dim(cm.data)[1]){
  vec<-cm.data[i,2:5]
  cm.data$prob[i]<-pmvnorm(lower=as.numeric(vec)-0.5,
                           upper=as.numeric(vec)+0.5,mean=pred.params.cm[,1],
                           sigma=cov.mat[6:9,6:9])[1]
}
cm.data$prob<-cm.data$prob/sum(cm.data$prob) # turn into percentage
# create array of years based on their similarity to prediction year
CMyear.sample<-sample(cm.data$year,5000,replace=TRUE, prob=cm.data$prob)
```

	% of sample
2000	99.5
2001	0.22
2010	0.02
2012	0.14
2014	0.12

Figure 5: Summary of center of mass sample

The resulting matrices are then saved as `.csv` to be used in the final simulation model.

2.5 Streamflow Simulation

The final irrigation season streamflow simulations are modeled in the `streamflow_simulation.R` script.

The original streamflow data, sampled volumes and centers of mass are imported to simulate the irrigation season hydrographs. This is done by selecting the timeseries of streamflow that corresponds with a given year from the center of mass sample and normalizing it by a volume from the multivariate distribution sample. This 'analog water year' approach effectively uses the linear models to estimate the most similar year in runoff timing, and normalizes that hydrograph based on the predicted volume estimates.

```
for(k in 1:ns){ # ns = number of simulations, in our case 5000
  # Simulate natural flow supply at the four gages and total diversions
  year<-cm.year[k,1] # year sample
  vol<-volumes[k,] # volume sample

  # select the streamflow timeseries that corresponds with the center of mass sample
  bwb<- bwb.wy[bwb.wy$wy == year, "bwb.q"][183:365]  # irrigation season
  # normalize the sampled hydrograph by the sampled volume
  bwb.flow.s[,k]<- bwb * vol/(sum(bwb)*1.98)
  # 1.98 is the conversion from cfs to ac-ft, (cfs) * (ac-ft/ac-ft)
```

Prediction intervals are calculated from the relevant quantiles from the simulation results

```
pred.int<-function(location){
  lo<-apply(location,1,quantile,0.05, na.rm=TRUE)
  hi<-apply(location,1,quantile,0.95, na.rm=TRUE)
  meanQ<-apply(location,1,mean, na.rm=TRUE)

  return(cbind(lo, hi, meanQ))
}
```

The following figure is an example model output figure for each basin, the average simulated hydrograph (blue), the prediction interval (shaded grey), and the actual hydrograph (green) for 2019.

```
#knitr::include_graphics(file.path(params$fig_dir_mo, "BWB_Simulation.png"))
```

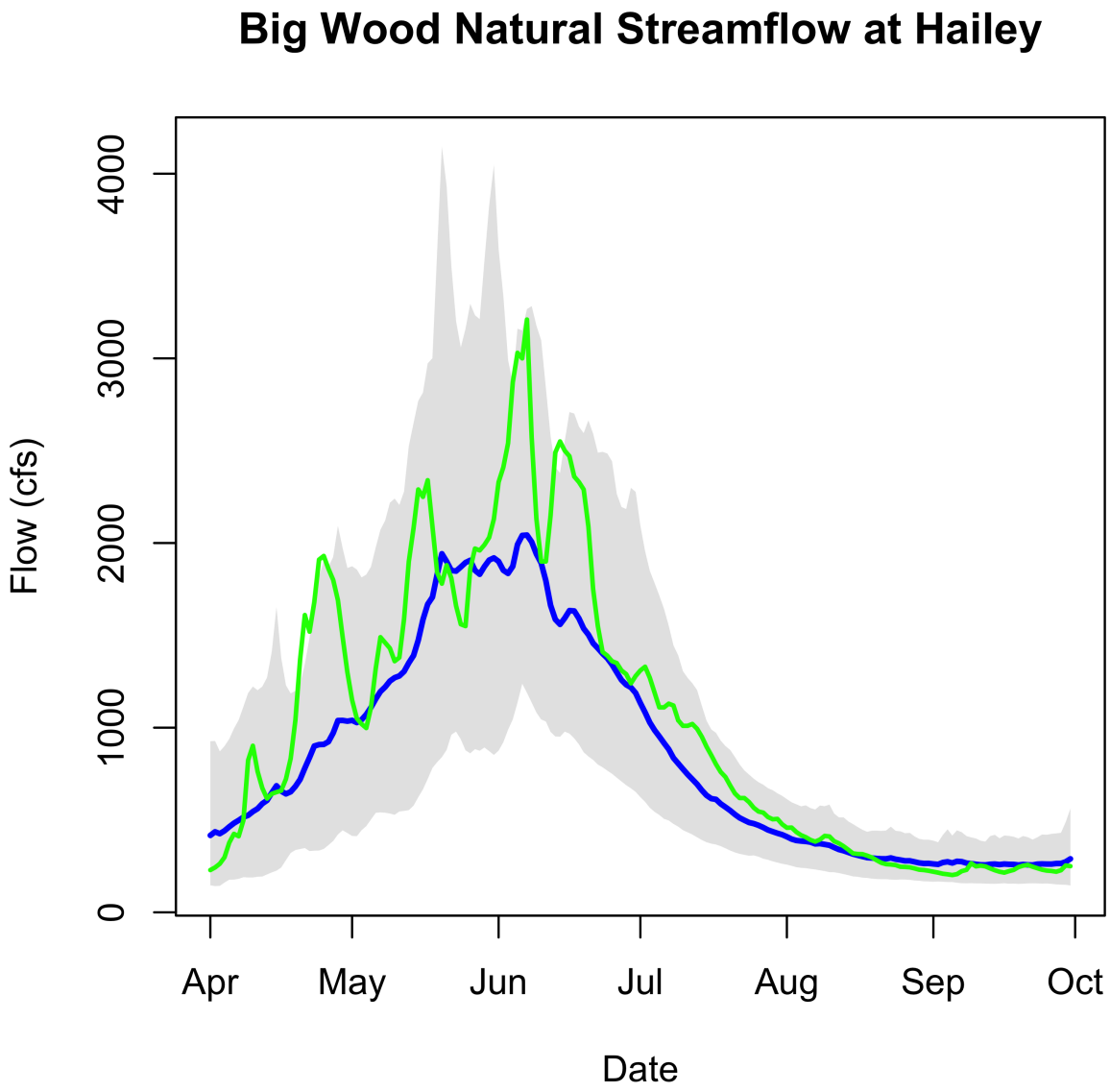


Figure 6: Simulated flows on the Big Wood River at Hailey

3. Overview of modeling results

While many of the individual linear regression models have strong fits, the compounding effects of multiple predictions make for large uncertainty windows. This is both valuable (ensures appropriate uncertainty in modeling results), and challenging for desired use of the model (narrower prediction windows for crop /

irrigation planning). This suite of models (1 temperature, 16 streamflow, 9 curtailment) and the automated process lays a strong groundwork for continued model development to fulfill the needs of the WRWC. Continued model development should focus on the components of the hydrograph that are most valuable to the members, namely the hydrograph recession. Ideally those modeling results would be available by April, but the largest uncertainty in the current models are spring weather (e.g. once we've reached the peak of the hydrograph, predicting the recession is relatively straight forward, but by that time it is later than needed by users.) Maintaining focus on the desired use of the model will be critical in prioritizing the next steps.

4. Recommendations

- continue to update diversion data annually
- Evaluate alternative methods for estimation of curtailment dates using all water rights data
- Evaluate use of available groundwater data (there may not be enough for this to be viable)
- Incorporate additional model variables (e.g. last years streamflow volume)
- Explore snow covered extent modeling
- Incorporate prediction data from the National Operational Hydrologic Remote Sensing Center, and other regional forecasting centers (summer temperatures in particular)
- Conduct a sensitivity analysis on prediction results to identify the most sensitive predictor variables, particularly on streamflow recession
- Incorporate downstream water rights for Silver Creek and GW diversions

5. Citations

(2008) Bayesian Information Criteria. In: Information Criteria and Statistical Modeling. Springer Series in Statistics. Springer, New York, NY. https://doi.org/10.1007/978-0-387-71887-3_9