ESM 232 - Hwk8

Julia Dagum and Mauricio Collado

5/22/2021

1. Objective

Our objective is to develop an environmental performance metric for the runs of different hydrologic models. In this case, we value the model's performance if it can predict the total monthly streamflow of the Sierra Watershed between October and May. We selected these months because ENSO events usually took place between October and February, followed by a recovery of the streamflow around April. In this context, we built a function to evaluate the correlation between the observed total monthly streamflow and the forecast. We use the total monthly streamflow as a proxy of water stress.

Many experts may argue that accurate estimates are more relevant for the summer season. As a response, it is easy to adapt our function to evaluate other seasons of the year.

2. Performance metric

Our performance metric calculates the correlation between the total monthly streamflow of the observed and predicted values. It prioritizes the accuracy between October and May, where ENSO events occur, followed by a streamflow recovery around April.

```
#' monthly total flow metrics
#'
#' Compute correlation between the monthly totals observation and model
#' @param  m  model estimates
#' @param  o  observations
#' @param  month month
#' @param  day day
#' @param  year year
#' @param  start start month for evaluation, the default value is 10 (October)
#' @param  end start month for evaluation, the default value is 5 (May)
#' @return monthly_tot_cor


check_sel_tot_mon = function(m, o, month, day, year, wy, start=10, end=5) {

  flow = cbind.data.frame(m, o, month, day, year,wy)
  # first lets get minimum yearly values

  tmp = flow %>%
    group_by(month, wy) %>% # group by month and year
    filter(month>=start | month<=end) %>% # OR for relevant months!!!
    summarize(tot_o=sum(o), #total flow for obervations
              tot_m=sum(m))

    sel_mon_tot_cor = cor(tmp$tot_m, tmp$tot_o)

  return(sel_mon_tot_cor=sel_mon_tot_cor)
}
```

3. Data format

It is necessary to format all the estimations by day, month, year, and wy (hydrological cycle). We worked with the model runs of sagerm and the dates and observations of sager datasets in this case.

```
# read the run dataset
msage = read.table(here("data","sagerm.txt"),
                   header=T)

# read the date and observation dataset
sage = read.table(here("data","sager.txt"),
                   header=T)
# bring date format
sage = sage %>%
  mutate(date=make_date(year=year, month=month, day=day))

# use the starting from sage and apply to msage
msage$date = sage$date
msage$month = sage$month
msage$year = sage$year
msage$day = sage$day
msage$wy = sage$wy

# apply also for observations
msage$obs = sage$obs

# turn all the columns of different outputs into a single column identified by "run"
msagel = msage %>% gather(key="run",
                          value="streamflow",
                          -date, -month, -day, -year, -wy, -obs)
```

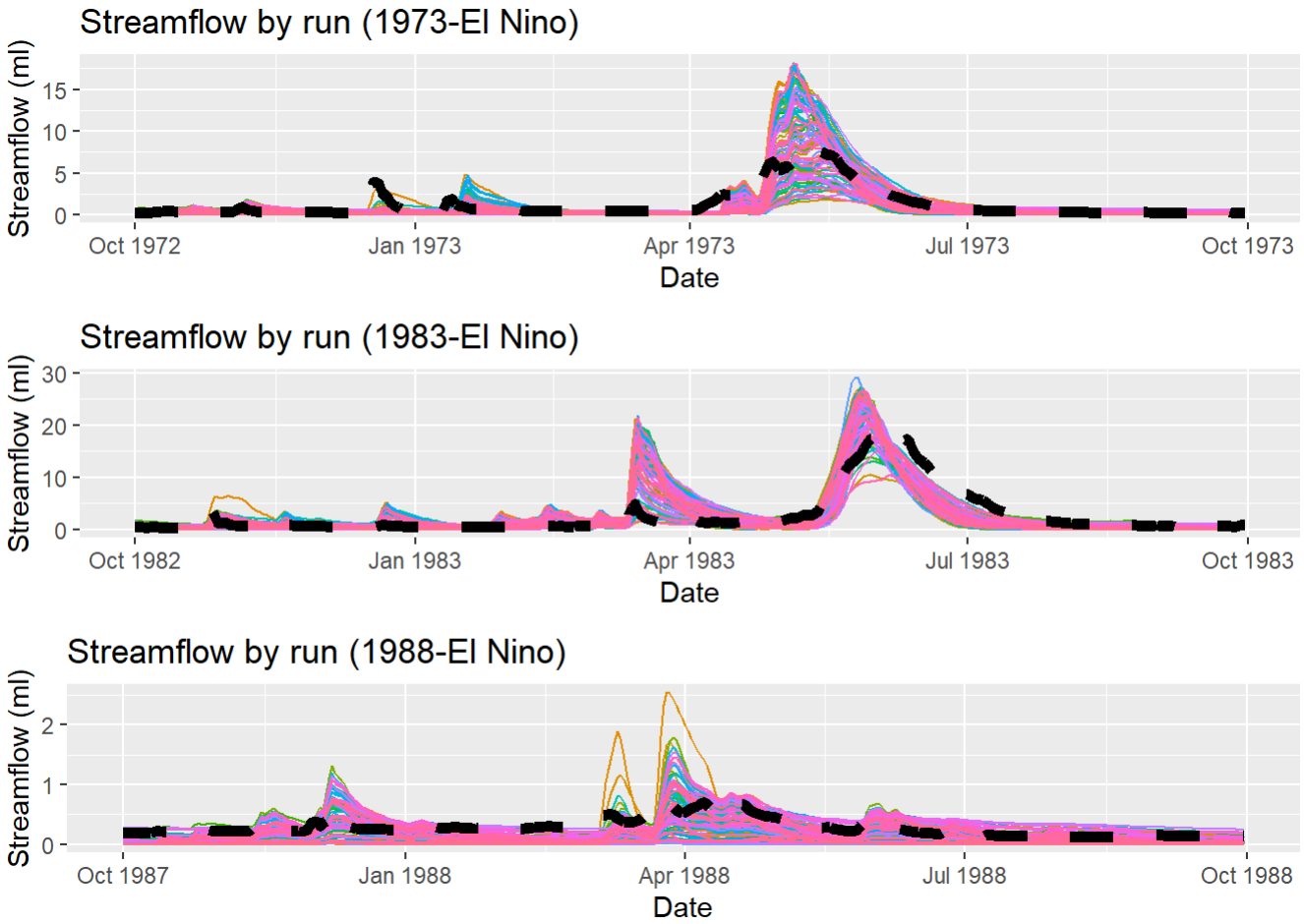
4. Selection of the analysis period

There is evidence that major ENSO events influence streamflow from Sierra Watershed. Consequently, we pick strong El Nino years (1972-73. 1982-1983, and 1987-88) to reference the altered streamflow.

```
# plot water

#1973
n1=ggplot(subset(msagel, wy == 1973), aes(as.Date(date), streamflow, col=run))+
  geom_line()+
  theme(legend.position = "none")+
  geom_line(aes(as.Date(date), obs), size=2, col="black", linetype=2)+
  labs(title="Streamflow by run (1973-El Nino)",
       y="Streamflow (ml)",
       x="Date")

#1983
n2=ggplot(subset(msagel, wy == 1983), aes(as.Date(date), streamflow, col=run))+
  geom_line()+
  theme(legend.position = "none")+
  geom_line(aes(as.Date(date), obs), size=2, col="black", linetype=2)+
  labs(title="Streamflow by run (1983-El Nino)",
       y="Streamflow (ml)",
       x="Date")

#1988
n3=ggplot(subset(msagel, wy == 1988), aes(as.Date(date), streamflow, col=run))+
  geom_line()+
  theme(legend.position = "none")+
  geom_line(aes(as.Date(date), obs), size=2, col="black", linetype=2)+
  labs(title="Streamflow by run (1988-El Nino)",
       y="Streamflow (ml)",
       x="Date")
```

The El Nino events happened between October (of the previous year) and February. During this period, we observe low streamflow, followed by an increase in April/May. The **model accuracy** will consider the El Nino period (Oct-Feb) and the streamflow recovery (March-April-May).

```
ggarrange(n1, n2, n3,
          ncol = 1, nrow = 3)
```



###

5.Performance metric analysis

Our first step is to run our function.

```
# run our monthly correlation metric
res = msage %>%
  select(-date, -month, -day, -year, -wy, -obs ) %>%
  map_dbl(~check_sel_tot_mon(o=msage$obs,
                             month=msage$month,
                             day=msage$day,
                             year=msage$year,
                             wy=msage$wy, m=.x))


# keep results
# naming runs
simnames = names(msage %>%
                   select(-date, -month, -day,-year,-wy, -obs))

# keep in dataframe
results = cbind.data.frame(simnames=simnames, moncorr=res)

# acceptable values
summary(results)
```

```
##    simnames           moncorr
##  Length:101         Min.   :0.7677
##  Class :character   1st Qu.:0.8337
##  Mode  :character   Median :0.8516
##                     Mean   :0.8482
##                     3rd Qu.:0.8657
##                     Max.   :0.8849
```

```
# The results are acceptable, all correlations are positive.

# see results
head(results)
```

```
##        simnames   moncorr
## V99.1     V99.1 0.8723346
## V100.1   V100.1 0.8663436
## V101       V101 0.8212585
## V102       V102 0.8554122
## V103       V103 0.8556660
## V104       V104 0.7993128
```

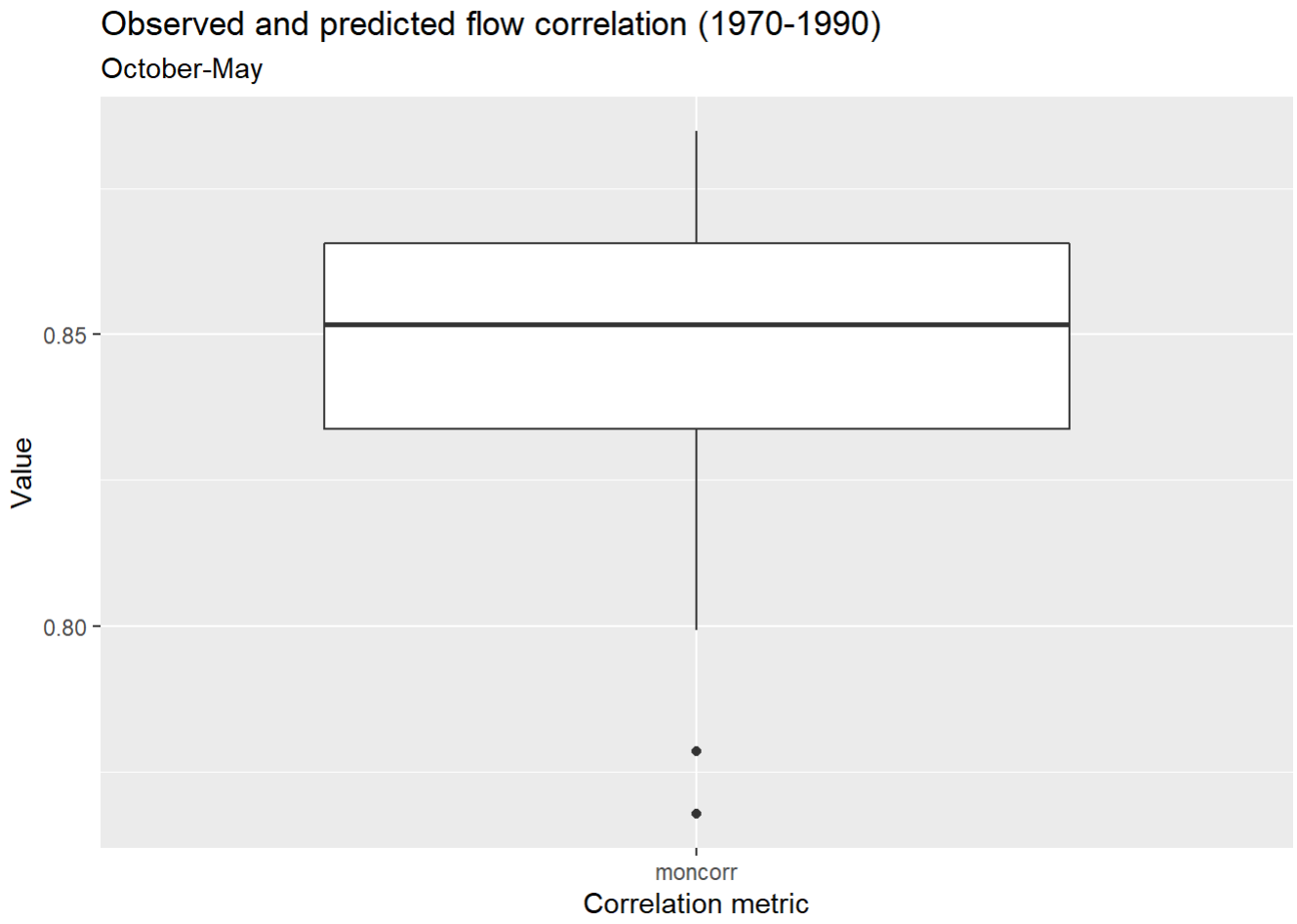
```
# keep best correlation
results[which.max(results$moncorr),]
```

```
##      simnames  moncorr
## V181     V181 0.884857
```

```
bestcorr <- row.names(results[which.max(results$moncorr),])
bestcorr_value <- round(max(results$moncorr), 6)
```

The run the provides the highest correlation is **V181** (0.884857). Our boxplot graph show that runs, on average, have a correlation of 0.85 with the observations. The run V181 is above average.

```
# graph range of performance measures
resultsl = results %>%
  gather(key="metric",value="value", -simnames)

# boxplot
ggplot(resultsl, aes(metric, value))+
  geom_boxplot() +
  labs(title="Observed and predicted flow correlation (1970-1990)",
       subtitle="October-May",
       y="Value",
       x="Correlation metric")
```


Observed and predicted flow correlation (1970-1990)
October-May
0.85
0.80
Value
moncorr
Correlation metric