Monthly data GAM and data preparation

Willem Vervoort, Michaela Dolk & Floris van Ogtrop

2017-05-18

```
# root dir
knitr::opts_knit$set(root.dir = "C:/Users/rver4657/ownCloud/Virtual Experiments/VirtExp")
knitr::opts_chunk$set(echo = TRUE)
# LOAD REQUIRED PACKAGES # #####
library(pander)
library(tidyverse)
library(zoo)
library(ggplot2)
library(hydromad)
library(Kendall)
library(mgcv)
library(doParallel)
library(foreach)
library(xts)
```

This rmarkdown document and the resulting pdf are stored on github. All directories (apart from the root working directory) refer to the directories in this repository.

Introduction

This document is related to the manuscript "Disentangling climate change trends in Australian streamflow" (vervoort et al.), submitted to Journal of Hydrology. This is the eightth part of the series that reruns the Mann Kendall and GAMM analysis on the monthly station data and creates monthly output for running the numerical modelling with monthly data.

So the steps that we need to do are:

- Aggregate the station data to monthly;
- run the Mann Kendall and GAMM analysis;
- write the monthly data as a datafile to be read in on the HPC modelling.

However, the monthly Mann-Kendall analysis has already been run in the 2nd part of the series (2.MannKendallTest.pdf), so we only need to run the GAMM analysis

Aggregate station data to monthly

```
load("data/ClimCh_project_MD.Rdata")


flow_zoo_m <- aggregate(flow_zoo,as.yearmon,sum,na.rm=T)
rain_zoo_m <- aggregate(rain_zoo,as.yearmon,sum,na.rm=T)
maxT_zoo_m <- aggregate(maxT_zoo,as.yearmon,sum,na.rm=T)

flow_rain_maxT_monthly <- melt(as.data.frame(rain_zoo_m))
```

```
## Using  as id variables
```

```r
colnames(flow_rain_maxT_monthly) <- c("Station","Rain")
flow_rain_maxT_monthly$Station <-
  substr(flow_rain_maxT_monthly$Station,1,4)

flow_rain_maxT_monthly$Flow <- melt(as.data.frame(flow_zoo_m))[,2]
```

```
## Using  as id variables
```

```r
flow_rain_maxT_monthly$MaxT <- melt(as.data.frame(maxT_zoo_m))[,2]
```

```
## Using  as id variables
```

```r
rm(list=c("flow_rain_maxT_weekly","GridRainAllDataout",
     "weekGridRainAllDataout",
     "flow_zoo","maxT_zoo", "rain_zoo"))
```

Run the GAMM analysis on the monthly data

Model 1 Only flow and trend

The first 2 models are actually not generalised additive mixed models (GAM) as the models only analyse a linear trend. To match the GAM analysis, we used generalised least squares (`gls()`) in R. This still allows correlated errors to be analysed

```r
# run the gls model on flowtrend only
cl <- makeCluster(4) # create a cluster with 4 cores
registerDoParallel(cl) # register the cluster
# use a foreach loop to calibrate
Store2 <- foreach(i = 1:length(Stations[,1]),
                 .packages="mgcv") %dopar% {
  gamm.data <- subset(flow_rain_maxT_monthly,
                      flow_rain_maxT_monthly$Station == Stations[i,1])
  gamm.data$trend <- 1:nrow(gamm.data)
  gam_TrendOnly <- gls(log(Flow +1)~trend, correlation= corCAR1(),
       data=gamm.data)
  out <- list(model = gam_TrendOnly,
             results = data.frame(Station=Stations[i,1],
             t(summary(gam_TrendOnly)$tTable[2,c(1,4)]),
                  AIC=summary(gam_TrendOnly)$AIC))
  out
}
stopCluster(cl)

par(mfrow=c(5,3),mar=c(2,2,2,2))
for (i in seq_along(Stations[,1])) {
  res <- residuals(Store2[[i]]$model)
  plot(res, main=Stations[i,1], cex.main=0.7,
       ylab="normalised residuals",xlab="")
  n <- length(res)
  abline(lsfit(1:n, res), col="red")
}
```

```r
# store results
storedir <- "c:/users/rver4657/owncloud/virtual experiments"
save(Store2,file=paste(storedir,
                "projectdata/Store2_MonthlyTrendOnly.RData",
                sep="/"))
output <- do.call(rbind, lapply(1:length(Store2), function(i) rbind(Store2[[i]][[2]])))
pander(output, caption="Mixed model results for analysis of trend in monthly flow only")
```

Table 1: Mixed model results for analysis of trend in monthly flow only

Station	Value	p.value	AIC
COTT	-0.001254	0.1664	886.2
RUTH	-0.003158	7.985e-07	1166
CORA	-0.001459	0.02268	1560
ELIZ	-0.0004399	0.7958	1690
COCH	4.564e-05	0.9514	982
COEN	-0.0001306	0.9346	1601
SCOT	-0.0005103	0.6448	1172
HELL	-0.000645	0.3894	1031
NIVE	-0.0002345	0.8016	1348
MURR	-0.0009173	0.1193	235.2
SOUT	-0.0006692	0.1052	590.4
YARR	-0.001043	0.1174	886.8
DOMB	-0.0009748	0.547	1366

```r
rm(Store2)
```

Model 2 trend in rain

Similar to the flow data, this analysis uses `gls()` to run the linear mixed model to test for a trend in the data and compare to the Mann-Kendall results

```r
# create an empty list
# and an empty dataframe to store results
# run the gls model on flowtrend only
cl <- makeCluster(4) # create a cluster with 4 cores
registerDoParallel(cl) # register the cluster
# use a foreach loop to calibrate
Store_Rain <- foreach(i = 1:length(Stations[,1]),
                 .packages="mgcv") %dopar% {
  gamm.data <- subset(flow_rain_maxT_monthly,
                      flow_rain_maxT_monthly$Station == Stations[i,1])
  gamm.data$trend <- 1:nrow(gamm.data)
  gam_TrendR <- gls(log(Rain + 1)~trend, correlation= corCAR1(),
      data=na.omit(gamm.data))
  out <- list(model = gam_TrendR,
              results = data.frame(Station=Stations[i,1],
                     t(summary(gam_TrendR)$tTable[2,c(1,4)]),
                              AIC=summary(gam_TrendR)$AIC))
  out
}
```

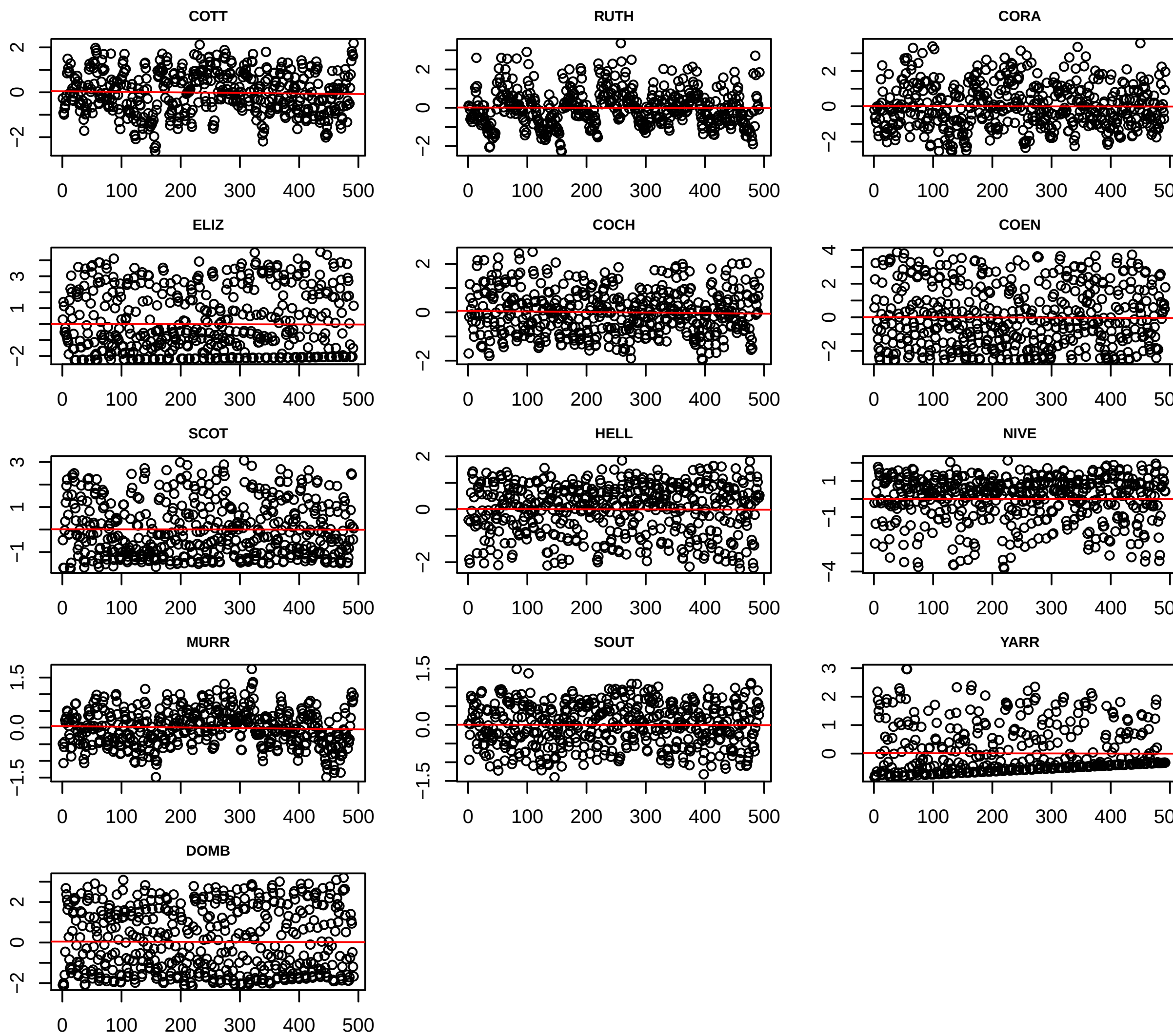

Figure 1: Residuals of linear mixed model analysis for trend in monthly flow only

```r
stopCluster(cl)


par(mfrow=c(5,3),mar=c(2,2,2,2))
for (i in seq_along(Stations[,1])) {
  res <- residuals(Store_Rain[[i]]$model)
  plot(res, main=Stations[i,1], cex.main=0.5,
       ylab="normalised residuals",xlab="")
  n <- length(res)
  abline(lsfit(1:n, res), col="red")
}

# store results
save(Store_Rain,file=paste(storedir,
                  "projectdata/StoreRain_MonthlyTrend.RData",
                    sep="/"))
output <- do.call(rbind, lapply(1:length(Store_Rain),
                     function(i) rbind(Store_Rain[[i]][[2]])))
pander(output, caption="Mixed model results for analysis of trend in Monthly Station Rainfall")
```

Table 2: Mixed model results for analysis of trend in Monthly Station Rainfall

Station	Value	p.value	AIC
COTT	0.001877	0.1666	1677
RUTH	-0.0005554	0.2094	1554
CORA	2.369e-05	0.9442	1368
ELIZ	0.003226	0.08427	1976
COCH	-0.0002598	0.6802	1618
COEN	-0.0001202	0.9319	1944
SCOT	-0.0002076	0.6759	1343
HELL	-0.0004494	0.05605	1006
NIVE	-0.0005041	0.115	1197
MURR	-0.0002322	0.482	1217
SOUT	-0.000529	0.2159	1234
YARR	-0.0003523	0.6952	1668
DOMB	7.281e-06	0.992	1398

```r
rm(Store_Rain)
```

Model 3 GAMM with rainfall

This model analyses flow as a function of rainfall only. This is therefore an analysis of the rainfall runoff coefficient, taking into account a possible time trend in the data. If the trend in this analysis is significant, then this is a measure of how the rainfall runoff coefficient has changed over time.

```r
# Gamm model with flow and rain
cl <- makeCluster(4) # create a cluster with 4 cores
registerDoParallel(cl) # register the cluster
# use a foreach loop to calibrate
Store_FwR <- foreach(i = 1:length(Stations[,1]),
                 .packages="mgcv") %dopar% {
```

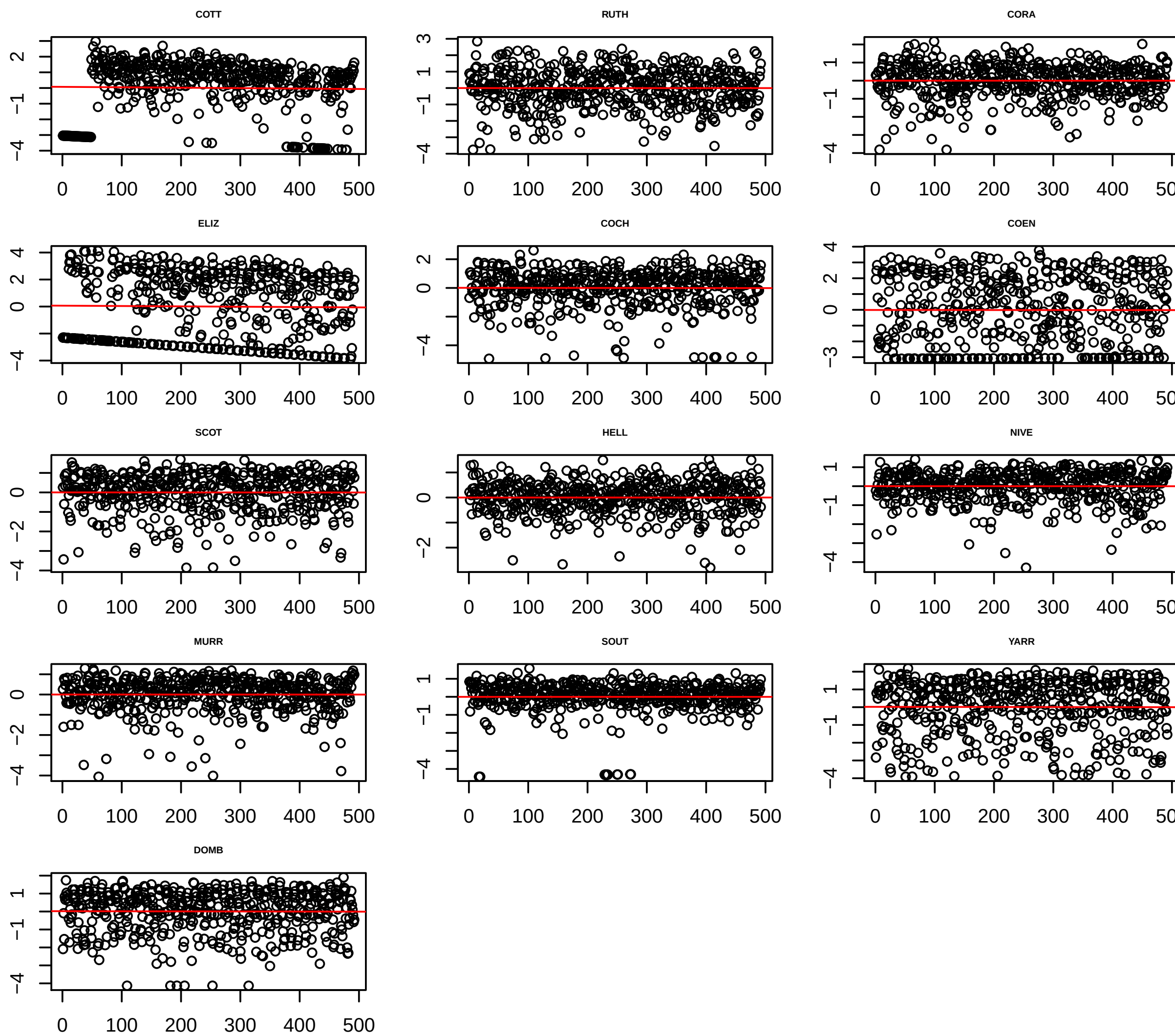

Figure 2: Residuals of linear mixed model analysis for trend in monthly station rainfall data

```
  gamm.data <- subset(flow_rain_maxT_monthly,
                      flow_rain_maxT_monthly$Station == Stations[i,1])
  gamm.data$trend <- 1:nrow(gamm.data)
  gam_TrendFlow_withR <- gamm(log(Flow +1)~s(Rain) + trend,
                               correlation= corCAR1(), data=gamm.data)
  out <- list(model = gam_TrendFlow_withR,
              results = data.frame(Station=Stations[i,1],
                           t(summary(gam_TrendFlow_withR$lme)$tTable[2,c(1,5)]),
                                  AIC=summary(gam_TrendFlow_withR$lme)$AIC))
  out
   }
stopCluster(cl)

par(mfrow=c(5,3), mar=c(2,2,2,2))
for (i in seq_along(Stations[,1])) {
  res <- residuals(Store_FwR[[i]]$model$lme)
  plot(res, main=Stations[i,1], cex.main=0.5,
       ylab="normalised residuals",xlab="")
  n <- length(res)
  abline(lsfit(1:n, res), col="red")
}

# store results
save(Store_FwR,
     file=paste(storedir,
              "projectdata/StoreFwR_monthlyTrend.RData",
              sep="/"))
output <- do.call(rbind, lapply(1:length(Store_FwR),
                                function(i) rbind(Store_FwR[[i]][[2]])))
pander(output, caption="Mixed model results for analysis of trend in monthly flow data taking into acco
```

Table 3: Mixed model results for analysis of trend in monthly flow data taking into account Rainfall

Station	Value	p.value	AIC
COTT	-0.001423	0.07129	719.5
RUTH	-0.002703	1.667e-05	684.9
CORA	-0.001096	0.01713	1169
ELIZ	-0.001192	0.314	1491
COCH	2.924e-05	0.9516	568.2
COEN	9.382e-05	0.9339	1205
SCOT	-0.0004628	0.5227	885.1
HELL	-0.0004809	0.4627	924
NIVE	-0.0001456	0.854	1299
MURR	-0.0008686	0.09077	49.24
SOUT	-0.0003789	0.2875	398.5
YARR	-0.0009496	0.03775	753
DOMB	-0.000986	0.3234	1085

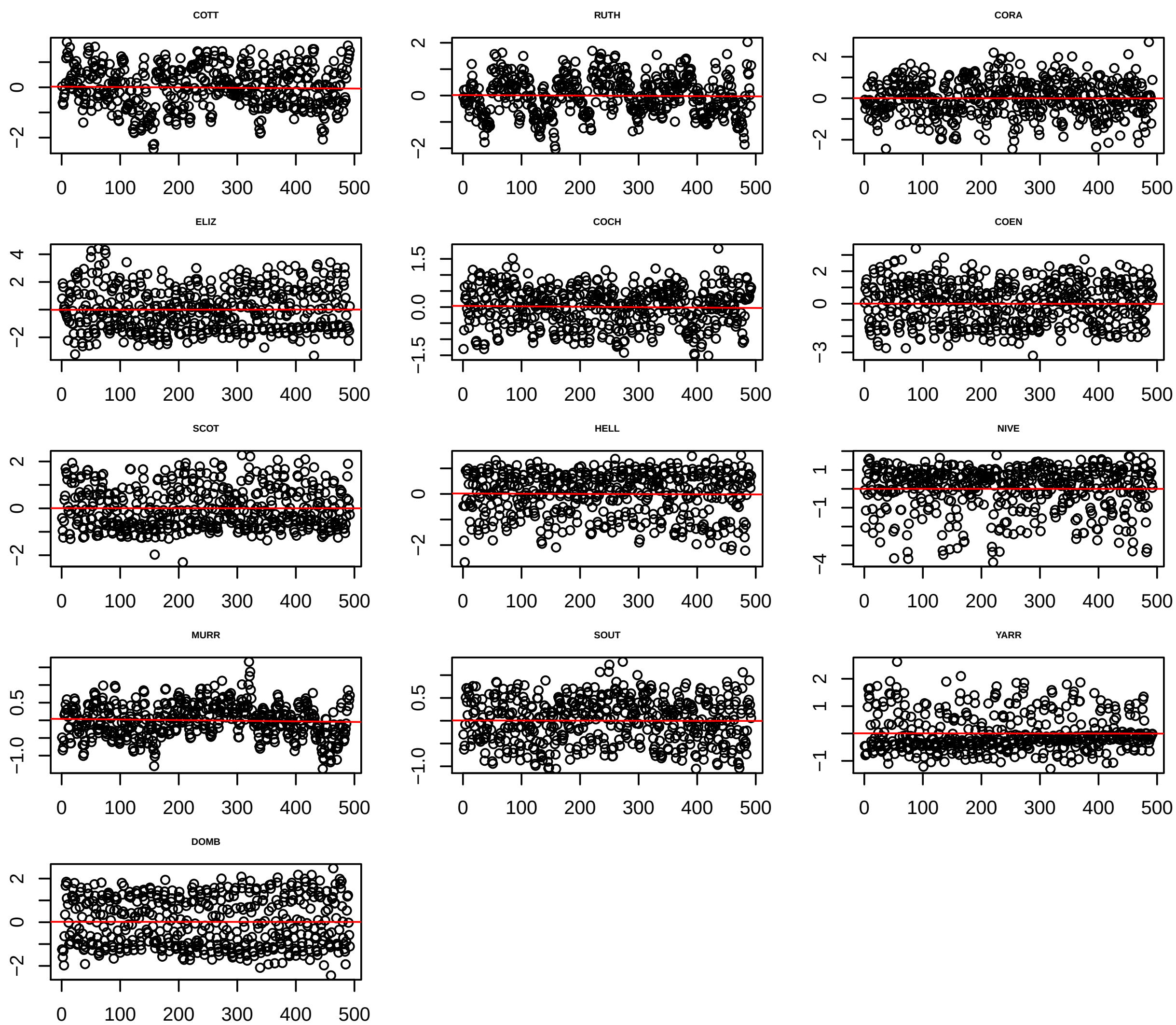

Figure 3: Residuals of GAMM analysis for trend in monthly flow data taking into consideration station rainfall data

```
rm(Store_FwR)
```

Model 4. GAMM with rain & s(rain,MaxT) and trend

This model analyses flow as a function of rainfall and the interaction between rainfall and maximum temperature, which is conceptualised as the actual evapotranspiration. This is therefore an analysis of the rainfall runoff coefficient, taking into account the changes in evapotranspiration and a possible time trend in the data. If the trend in this analysis is significant, then this is a measure of how the rainfall runoff coefficient has changed over time.

```
# run the gamm model on rain, maxT and flow
cl <- makeCluster(4) # create a cluster with 4 cores
registerDoParallel(cl) # register the cluster
# use a foreach loop to calibrate
Store_FwRE <- foreach(i = 1:length(Stations[,1]),
                  .packages="mgcv") %dopar% {
  gamm.data <- subset(flow_rain_maxT_monthly,
                      flow_rain_maxT_monthly$Station == Stations[i,1])
  gamm.data$trend <- 1:nrow(gamm.data)
  gam_TrendFlow_withRandE <- gamm(log(Flow +1)~s(Rain) + s(Rain, MaxT) +
                                       trend, correlation= corCAR1(),
                                       data=gamm.data, control=list(niterEM=0))
  out <- list(model = gam_TrendFlow_withRandE,
        results = data.frame(Station=Stations[i,1],
                      t(summary(gam_TrendFlow_withRandE$lme)$tTable[2,c(1,5)]),
                            AIC=summary(gam_TrendFlow_withRandE$lme)$AIC))
  out
 }
stopCluster(cl)

par(mfrow=c(5,3), mar=c(2,2,2,2))
for (i in seq_along(Stations[,1])) {
  res <- residuals(Store_FwRE[[i]]$model$lme)
  plot(res, main=Stations[i,1], cex.main=0.5,
       ylab="normalised residuals",xlab="")
  n <- length(res)
  abline(lsfit(1:n, res), col="red")
}

# store results
save(Store_FwRE,
     file=paste(storedir,
                "projectdata/StoreFwRE_MonthlyTrend.RData",
                sep = "/"))
output <- do.call(rbind, lapply(1:length(Store_FwRE),
                             function(i) rbind(Store_FwRE[[i]][[2]])))
pander(output, caption="Mixed model results for the analysis of trend in monthly flow data taking into a
```

Table 4: Mixed model results for the analysis of trend in monthly flow data taking into account Rainfall and Evapotranspiratiion

Station	Value	p.value	AIC
COTT	-0.001289	0.07837	686.8

Station	Value	p.value	AIC
RUTH	-0.002629	5.736e-06	667.1
CORA	-0.001233	0.002365	1148
ELIZ	-0.001133	0.3205	1487
COCH	2.836e-05	0.9411	510.6
COEN	0.0004392	0.6453	1178
SCOT	-0.0005162	0.2847	763.9
HELL	-0.0004326	0.1339	731.3
NIVE	-2.717e-05	0.9556	1145
MURR	-0.000821	0.08388	-3.948
SOUT	-0.0003229	0.1996	313.1
YARR	-0.0008982	0.01756	714.7
DOMB	-0.0007776	0.1267	859.8

```
rm(Store_FwRE)
```

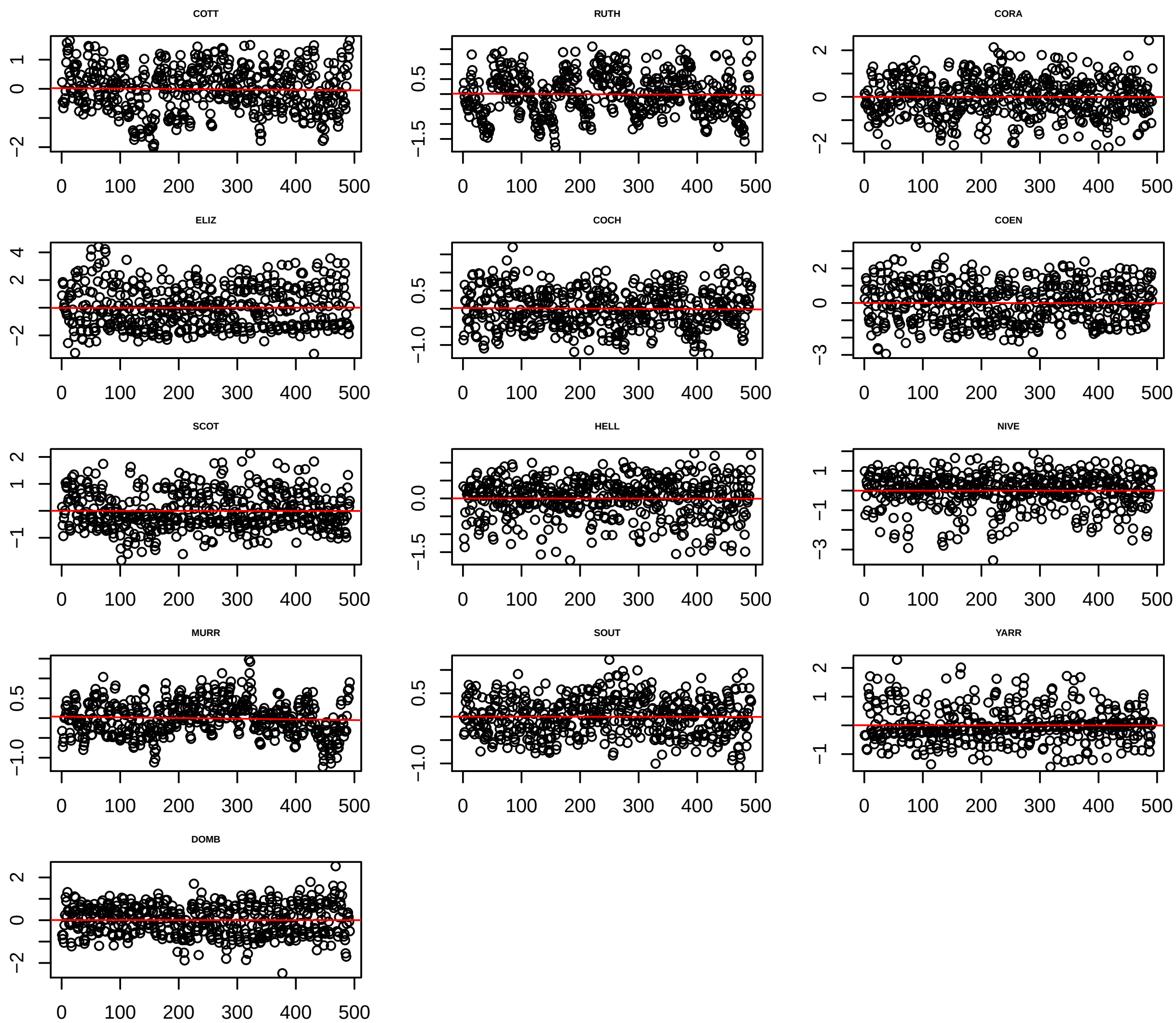

Figure 4: Residuals of GAMM analysis for trend in monthly flow data removing station rainfall and evapotranspiration effects

Model 5, same as model 4, but no trend and Mann Kendall on the residuals

This last model is to check the trend with GAMM analysis with the analysis using Mann-Kendall. So rather than incorporating a trend in the model, we analyse the residuals using Mann-Kendall for a trend. In this case we drop the plotting of the residuals.

```
# run the gamm model on rain, maxT and flow
cl <- makeCluster(4) # create a cluster with 4 cores
registerDoParallel(cl) # register the cluster
# use a foreach loop to calibrate
Store_FwRE2 <- foreach(i = 1:length(Stations[,1]),
                 .packages="mgcv") %dopar% {
  gamm.data <- subset(flow_rain_maxT_monthly,
                      flow_rain_maxT_monthly$Station == Stations[i,1])
  gam_Flow_withRandE <- gamm(log(Flow +1)~s(Rain) + s(Rain, MaxT) ,
                             correlation= corCAR1(), data=gamm.data,
                             control=list(niterEM=5))
  out <- list(model = gam_Flow_withRandE,
        results = data.frame(Station=Stations[i,1],
                         AIC=summary(gam_Flow_withRandE$lme)$AIC))
  out
}

stopCluster(cl)

# store results
save(Store_FwRE2,
     file=paste(storedir,
                "projectdata/StoreFwRE_Monthly.RData",
                sep="/"))
output <- do.call(rbind, lapply(1:length(Store_FwRE2),
                              function(i) rbind(Store_FwRE2[[i]][[2]])))
pander(output, caption="Mixed model results for the analysis of monthly flow data taking into account Ra
```

Table 5: Mixed model results for the analysis of monthly flow data taking into account Rainfall and Evapotranspiration

Station	AIC
COTT	687.8
RUTH	681
CORA	1155
ELIZ	1486
COCH	508.6
COEN	1176
SCOT	763
HELL	731.5
NIVE	1143
MURR	-3.229
SOUT	312.8
YARR	718.1
DOMB	860.1

Now do the Mann-Kendall analysis on the residuals

```
# do mann kendall on the residuals
resid_list <- vector("list", length=length(Stations[,1]))
for (i in seq_along(Stations[,1])) {
  resid_list[[i]] <- zoo(residuals(Store_FwRE2[[i]]$model$lme,
                type="normalized"),
          order.by=time(flow_zoo_m))
}
resid_df <- do.call(merge.zoo,resid_list)
names(resid_df) <- Stations[,1]
# Bootstrap
set.seed(10)
# now run a loop over the number of years (create 41 different sets)
# do Mann Kendall test on each resonstituted series
# ---------------------------
#  -------------------------
resid_temp <- as.data.frame(resid_df)
resid_temp$years <- format(time(resid_df),"%Y")
split_resid <- split(resid_temp[,1:13],resid_temp$years)


cl <- makeCluster(4) # create a cluster with 4 cores
registerDoParallel(cl) # register the cluster
# use a foreach loop to calibrate
MK_list <- foreach(i = 1:500,
                 .packages=c("Kendall", "xts")) %dopar% {
  # reorganise the list elements
  series <- sample(1:nyears(resid_df),nyears(resid_df))
  for (j in 1:length(series)) {
    if (j==1) {
      new_df <- as.data.frame(split_resid[[series[j]]])
    } else {
      new_df <- rbind(new_df,as.data.frame(split_resid[[series[j]]]))
    }
  }
  # run mann kendall on the columns and store the results
  mk_r <- apply(new_df,2,MannKendall)

  out <- do.call(cbind,mk_r)
  out
 }
stopCluster(cl)


MK_df <- do.call(rbind,MK_list)

pvalues <- subset(MK_df, rownames(MK_df)=="sl")
tau <- subset(MK_df, rownames(MK_df)=="tau")

sig_set <- list()

for (i in 1:ncol(pvalues)) {
  set <- data.frame(pvalue=as.numeric(pvalues[,i]),
                    tau=as.numeric(tau[,i]),
```

```r
                    catch=rep(colnames(MK_df)[i],nrow(tau)))
  sig_set[[i]] <- set[set$pvalue < 0.5,]
}

sig_set_a <- do.call(rbind,sig_set)
sig_set_a$type <- rep("bootstrap",nrow(sig_set_a))


MK_resid <- do.call(rbind,lapply(resid_list,MannKendall))


real_df <- data.frame(pvalue = as.numeric(MK_resid[,2]),
                      tau = as.numeric(MK_resid[,1]),
                      catch=Stations[,1],
                      type=rep("real",nrow(MK_resid)))
# A histogram of taus

hp <- ggplot(sig_set_a, aes(x=tau)) + geom_histogram(binwidth=0.03,colour="white")
# Histogram of significant tau's, divided by catch
# With panels that have the same scaling, but different range
# (and therefore different physical sizes)
hp <- hp + facet_wrap(~ catch,ncol=5)
# add a red point for the real slope from the data
p_value <- ifelse(real_df$pvalue<0.05,"< 0.05",">= 0.05")
hp <- hp + geom_point(data=real_df,aes(x=tau, y=0,colour=p_value),
                shape=16,size=5) +
  facet_wrap(~ catch,ncol=5)+ ggtitle("Residuals Streamflow after GAM") #+
hp <- hp + scale_colour_grey(start = 0, end = 0.6)
print(hp)
```

```r
pander(real_df, caption="Mann Kendall results for the Monthly GAMM residuals, ref Figure 8")
```

Table 6: Mann Kendall results for the Monthly GAMM residuals, ref Figure 8

pvalue	tau	catch	type
0.1223	-0.04661	COTT	real
2.057e-06	-0.1432	RUTH	real
0.001959	-0.0934	CORA	real
0.6888	-0.01209	ELIZ	real
0.7621	0.00914	COCH	real
0.5323	0.01884	COEN	real
0.1397	-0.04456	SCOT	real
0.1273	-0.046	HELL	real
0.9433	-0.002153	NIVE	real
0.1775	-0.04068	MURR	real
0.3491	-0.02825	SOUT	real
0.183	-0.04017	YARR	real
0.02024	-0.07004	DOMB	real

```r
rm(Store_FwRE2)
```

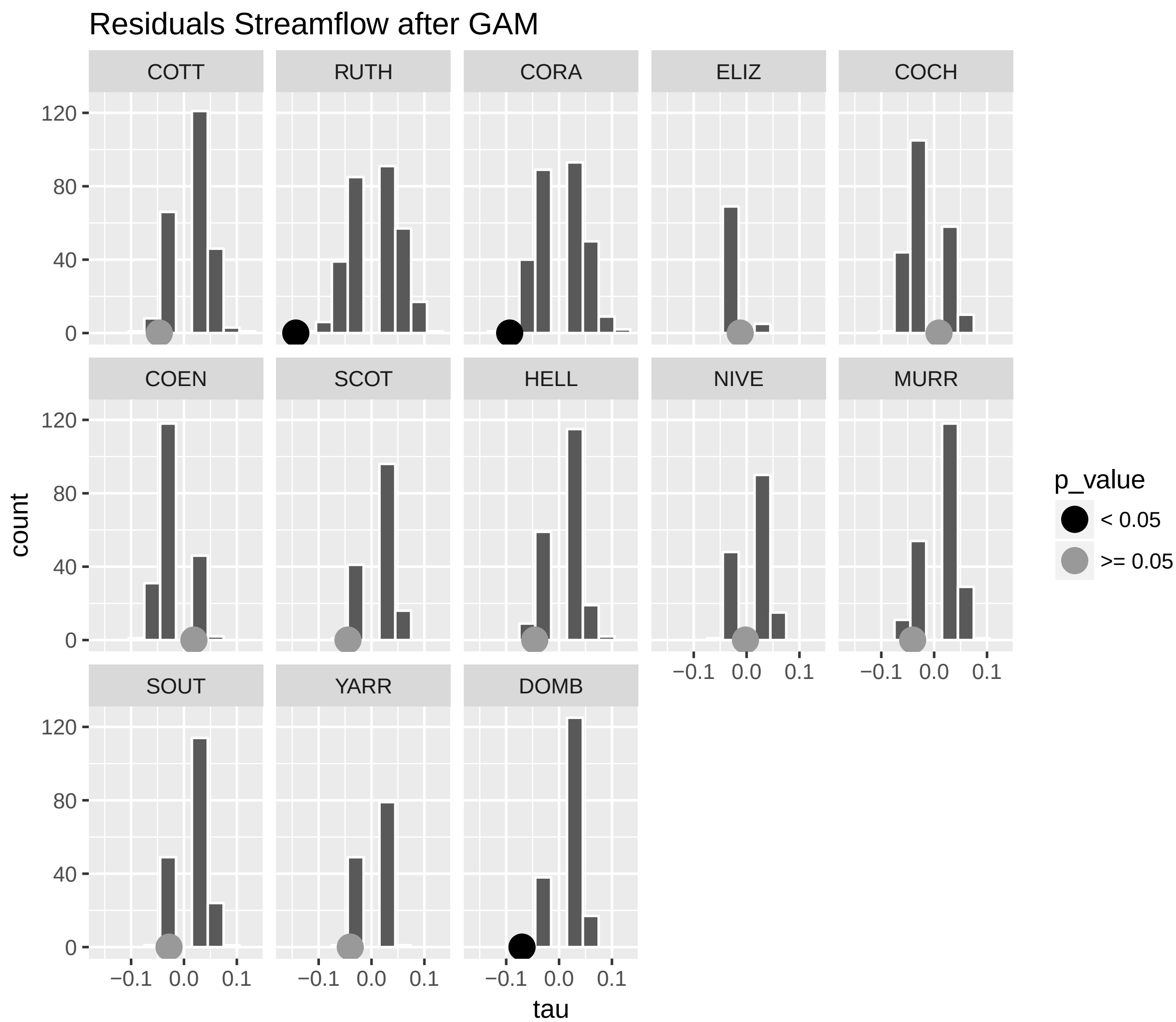

Figure 5: Mann Kendall analysis of the residuals of the monthly streamflow after GAM model with rainfall and a Evapotranspiration

Write the monthly data as a datafile to be read in on the HPC modelling

```
dataOut_Month <- list()

dataOut_Month[[1]] <- flow_zoo_m
dataOut_Month[[2]] <- rain_zoo_m
dataOut_Month[[3]] <- maxT_zoo_m
dataOut_Month[[4]] <- flow_rain_maxT_monthly

save(dataOut_Month,file="../projectData/MonthlyDataOut.Rdata")
```