Critique of Laufenberg et al. 2020

Stat 512:Final Project

Steve Huysman & Parker Levinson

Friday, May 5, 2023

Contents

1 Introduction

Whitebark pine (*Pinus albicaulis* Engelm.; abbreviation: WBP) is a conifer tree native to the mountains of the western United States and Canada. It inhabits subalpine areas where it can be found growing up to the tree line, often at a higher elevation than other tree species found in the same area. Whitebark pine is an early successional species that is often the first to establish after disturbance such as wildfire. It is a keystone species of subalpine environments where it plays important ecological roles such as providing food for wildlife such as Clark's Nutcracker and the threatened Grizzly Bear.

Due to threats from climate change, mountain pine beetle, and the invasive white pine blister rust, WBP has undergone a rapid and widespread decline. It was recently estimated that over half of all standing WBP in the United States are dead. This decline has lead to its recent listing as Threatened under the Endangered Species Act. Future climate projections indicate further deterioration of WBP's habitat. Strategies to conserve this species involve planting WBP seedlings for restoration of high-elevation forests. Successful plantings in the face of climate change require an understanding of the relationship between climate and seedling establishment and growth in this species. Climate can be investigate with a water balance model which incorporates the simultaneous availability of energy and water to plants in the form of measurements such as Actual Evapotranspiration (AET) and Climatic Water Deficit (CWD). Competition from other tree species also plays a role in seedling establishment and was investigated here.

Laufenberg et al. 2020 attempted to identify what climate variables were related to WBP seedling growth rates. Instead, we address a directed research question: How do growing season (April - October) actual evapotranspiration (AET) and growing season climatic water deficit (CWD) affect planted WBP seedling growth after accounting for location (planting unit and site), competition, and climatic variables? In addition, we are interested in the variation between planting sites.

2 Methods

2.1 Field Methods/Study design

Over the past 40 years, the US Forest Service and National Park Service has planted more than 1,500 acres of WBP in the GYE. This study used a hierarchical sampling design including 5 planting units (Appendix Figure 7) with a total of 29 planting sites across units. Whitebark seedlings were randomly sampled from the thousands of whitebark pine seedlings planted at each planting site. (See Figure 1 for sampling design.)

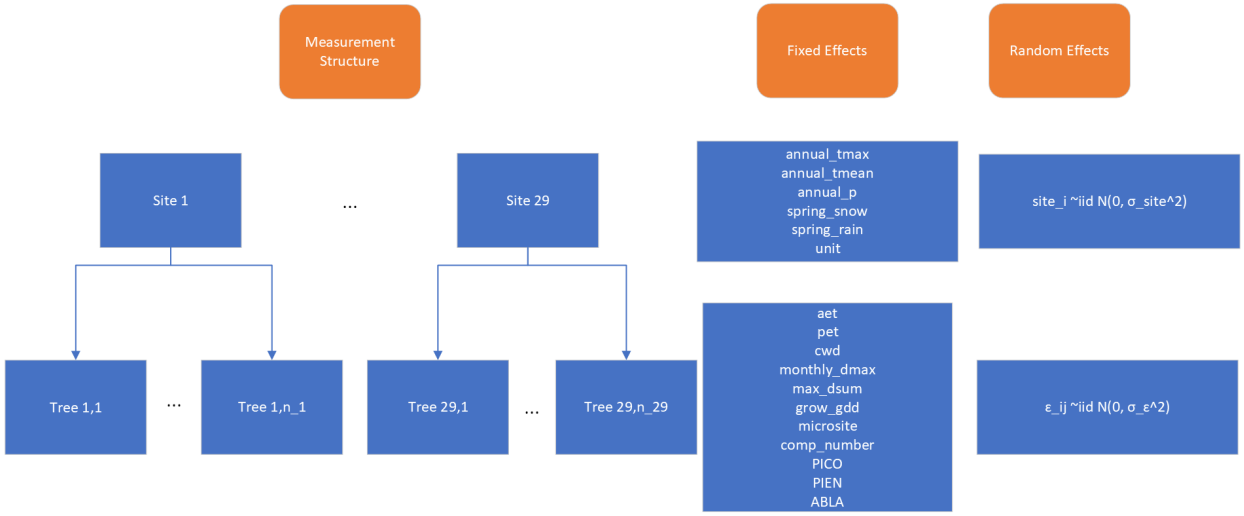


Figure 1: Hierarchical study design with fixed and random effects. Only variables addressed in our reserach question are included here.

The 29 sites sampled were unevenly spread across the 5 planting units, in an unbalanced study design. For each site, historical temperature and precipitation values were estimated using data from Daymet (Thornton et al. 2016), which were aggregated into seasonal and annual averages (annual_tmax, annual_tmean, spring_snow, spring_rain). A total of 1244 trees were sampled.

Seedlings within each sample site were then sampled for annual growth rate. A grid cell matrix of 10 meters x 10 meters was overlaid on the study site. A random starting point was decided and then every 20th grid cell from that was sampled, equating to sampling WBP in

2-15% of each site. Each seedling within that grid cell was digitally tagged, and Survey123 was used to collect field data. Seedlings were too small to measure growth rings via coring, so height was used as a proxy for growth rate. Specifically, growth rate was calculated as the change in height between the study year (2018) to the relative planted height when the seedling was first planted. This was divided by the number of years since planting minus 2.5 years to account for the period of time when seedlings sequester carbon instead of focusing on their own growth.

For each seedling, data was collected on the physical and topographic characteristics of the growing environment (point coordinates, elevation, aspect, and slope) as well as on the competition with other tree species (number of competitors, presence/absence of specific competitor species). Topographic data was used with the site-level climate data to model water balance variables at the tree-level (aet, pet, cwd). (For a full list of data collected at each location, see Figure 1

A variety of data were collected at different scales. The variables that are most pertinent to the research question are:

- *growth_rate* (*cm/year*): WBP seedling height is used as a proxy for growth rate. This is a continuous response variable.
- *AET* (*mm*): This is a continuous predictor variable for growing season actual evapotranspiration: the water loss through transpiration by plants on a site, given prevailing water availability. This measurement indicates the magnitude and duration of conditions favorable for plant growth on a site.
- *CWD* (*mm*): This is a continuous predictor variable for growing season climatic water deficit, which measures the evaporative water demand that is not met by the water supply at a site. It reflects the drought stress experienced by plants.

For the full list of variables that were collected, see appendix section 'Full list of variables analyzed in study'.

Data were provided by David Laufenberg, but the process of cleaning and structuring the data were not explicit in the paper. As such, we had to experiment to figure out how data were cleaned. Values for AET and CWD appeared to have been provided as an average for one month of the growing season (April - October: 7 months), as multiplying them by 7

resulted in the values reported for the growing season in Laufenberg et al. 2020. Thus, we multiplied the provided grow_dmean and grow_aetmean values by 7 to obtain values for the growing season sums for CWD and AET, respectively, which were used in our analysis.

2.2 Statistical Procedures Used

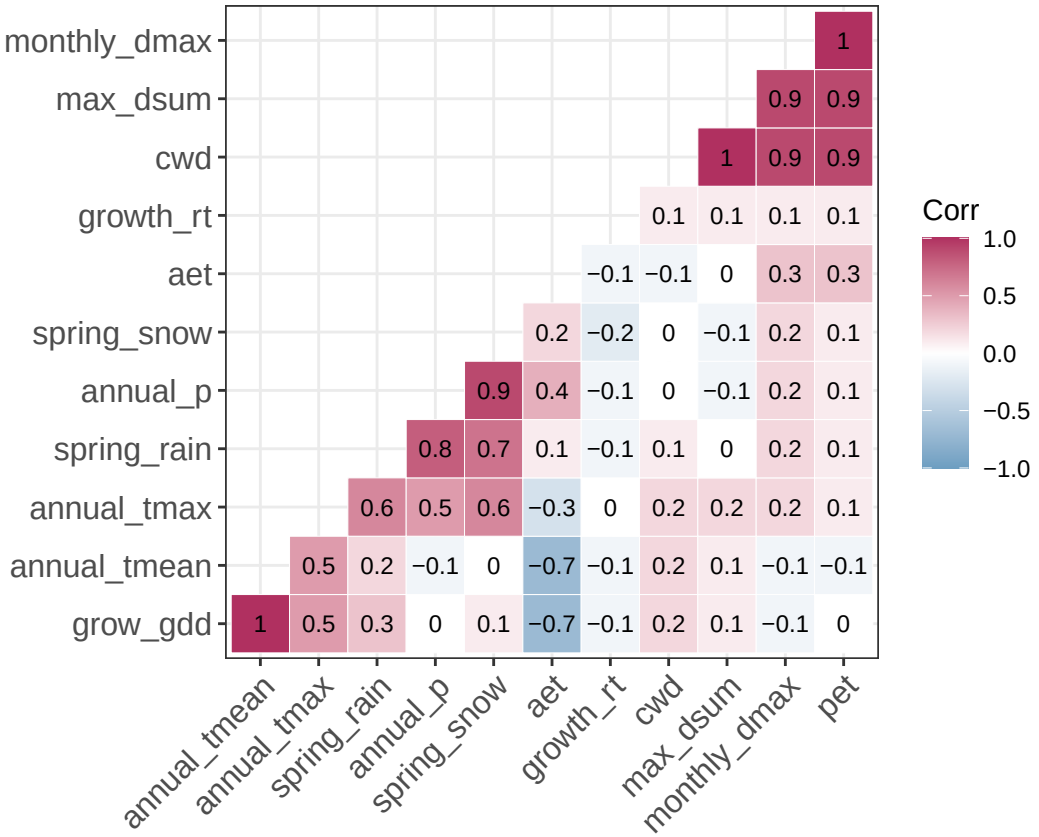


Figure 2: Correlation matrix of predictor variables. Darker squares indicate a higher degree of correlation. For variables that were at least 0.6 correlated, the most ecologically relevant one was selected.

Climate and water balance predictor variables were tested for colinearity with a cutoff of $r \geq \pm 0.6$ (Figure 2). A parsimonious list of variables was selected by choosing the more biologically relevant variable from pairs that exceeded this threshold. This list of variables was combined with competition and microsite covariates to include in our model selection process. These variables were assessed for linearity (Figure 3). We then looked at variation between sites (Figure 4), which confirmed the need to include a random effect of site.

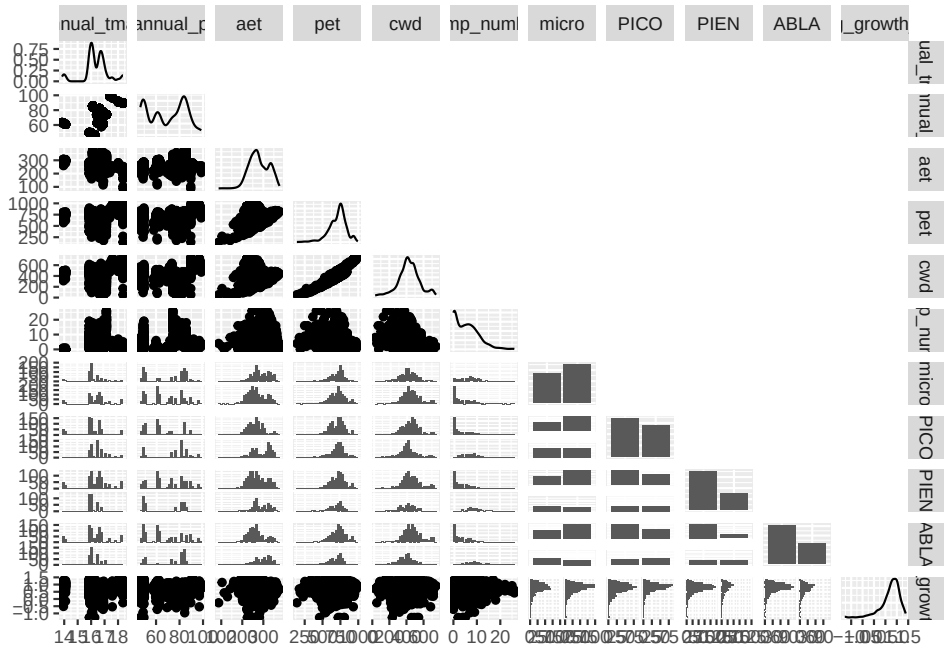


Figure 3: Raw data visualization of selected predictor variables

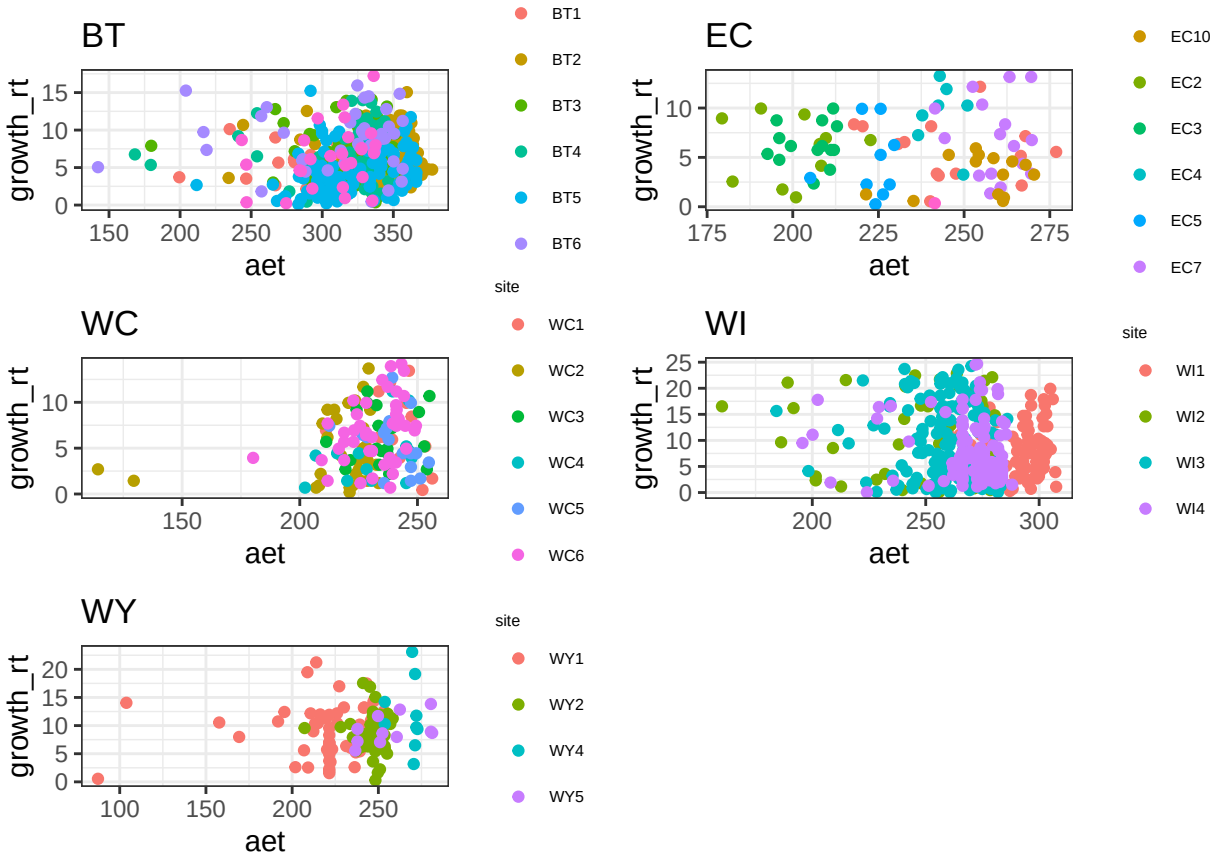


Figure 4: Growth rate compared with AET for each study site, by planting unit

Mixed effects models were used to investigate the relationship between WBP seedling growth and AET and CWD, after accounting for competition, other climatic variables, and planting location. Site was included as a random effect. Unit was included as a fixed effect varying at the site level, due a small sample size ($n_{unit} = 5$). Corrected AIC (AICc) was used to compare models. While the original study (Laufenberg et al. 2020) assessed cubic forms of AET, CWD, and comp_number in their model selection process, the biological reasoning behind this inclusion of higher-order terms and methodology for their assessment in model selection was unclear. Due to a lack of reasoning for their inclusion, we analyzed only first order terms in our model selection process here.

The theoretical full model is below:

$$
\begin{aligned}
log(growth\_rate)_{ij} &= \mu_{ij} + Site_i + \epsilon_{ij} \qquad (1)\\
Site_i &\overset{iid}{\sim} \mathcal{N}(0, \sigma^2_{Site})\\
\epsilon_{ij} &\overset{iid}{\sim} \mathcal{N}(0, \sigma^2_{error})\\
\mu_{ij} &= \beta_0 + \beta_1 * AET + \beta_2 * CWD + \beta_3 * I_{Unit=EC} + \beta_4 * I_{Unit=WC}\\
&\quad + \beta_5 * I_{Unit=WI} + \beta_6 * I_{Unit=WY} + ...\beta_p * climate\ variable_p
\end{aligned}
$$

$$
I_{unit=WC} = \begin{cases} 1, & \text{if } unit = West\ Centennial \\ 0, & \text{otherwise} \end{cases} \qquad I_{unit=EC} = \begin{cases} 1, & \text{if } unit = East\ Centennial \\ 0, & \text{otherwise} \end{cases}
$$

$$
I_{unit=WI} = \begin{cases} 1, & \text{if } unit = Wind\ River \\ 0, & \text{otherwise} \end{cases} \qquad I_{unit=WY} = \begin{cases} 1, & \text{if } unit = West\ Yellowstone \\ 0, & \text{otherwise} \end{cases}
$$

β_1 and β_2 will allow us to answer the research question of how AET and CWD affect

WBP seedling growth. $\beta_3 - \beta_6$ help get at the question of what variability exists in WBP seedling growth between planting units.

We ran the saturated model without interactions (Equation (1)), and then refined it using backwards selection from the *step*() function and an AIC cutoff of AIC<2.

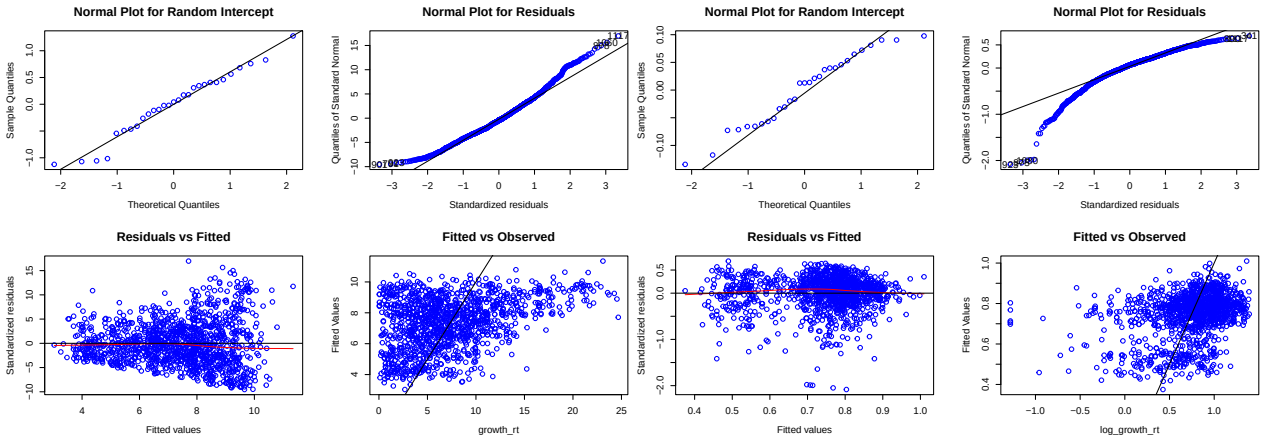


Figure 5: Diagnostic plots of saturated model before and after log transformation of growth rate

After running the model initially, we found that the residuals showed a funnel pattern (Figure 5), and we found that log transforming the response variable ($growth\_rate$) resulted in more constant variance. However, log transforming does result in a non-normal distribution of residuals, so we tried a variety of different functional transformations to both the response and predictor variables but were unable to reduce this violation. Ultimately, we decided that maintaining constant variance, especially with our large sample size, was more important so we proceeded with the log transformed response. Following the original study (Laufenberg et al. 2020), log transformations were performed with base 10. Measurements for individual seedlings are assumed to be independent due to spatial separation between seedlings and due to the fact that both site-to-site and unit-to-unit variation were incorporated into the model. There were no apparent violations of linearity from the raw data plot (Figure 3), and colinearity was addressed by removing highly correlated values. All analysis was done use R statistical software (v4.2.2; R Core Team 2021).

3 Results/Summary of Statistical Findings

Using backwards selection, the most parsimonious model included a random effect of site and fixed effects for unit, AET, and number of competitors (For full step selection process refer to code included in the appendix.). Although water deficit (CWD) was not included based on these model selection criteria (ΔAICc = 16.815), we included it because it directly helps us answer our research question. The select model had an AICc of 1148.38. The estimated model is below:

$$
\begin{aligned}
log(growth\_rate)_{ij} &= \hat{\mu}_{ij} + Site_i + \epsilon_{ij} \\
Site_i &\overset{iid}{\sim} \mathcal{N}(0, \hat{\sigma}^2_{Site}) \\
\epsilon_{ij} &\overset{iid}{\sim} \mathcal{N}(0, \hat{\sigma}^2_{error}) \\
\hat{\mu}_{ij} &= 0.22 + 0.0014 * AET - 0.000094 * CWD + 0.15 * I_{Unit=EC} + 0.14 * I_{Unit=WC} \\
&\quad + 0.18 * I_{Unit=WI} + 0.36 * I_{Unit=WY} + 0.0084 * comp\_number
\end{aligned}
$$

The final model suggests that WBP seedling growth is most correlated with the annual evapotranspiration and the number of competitors around the seedling. For a one mm change in growing season AET, there is a multiplicative change in the median seedling growth rate by 1.0033351 cm/year (95% CI: 1.000 to 1.007 cm/year, profile-likelihood CI) after accounting for CWD, number of competitors, unit, and site-to-site variability in seedling log growth rates. For a one mm change in CWD, there is a multiplicative change in the median seedling growth rate by 0.9997825 cm/year (95% CI: 0.999 to 1.001 cm/year, profile-likelihood CI), after accounting for AET, number of competitors, unit, and site-to-site variability in seedling log growth rates.

There is a difference in growth rate by unit, which can be visualized in the effects plot (Figure 6). After accounting for AET, water deficit, and number of competitors, the estimated correlation between any two seedlings in the same site is $\hat{ICC} = 0.217$. We found strong

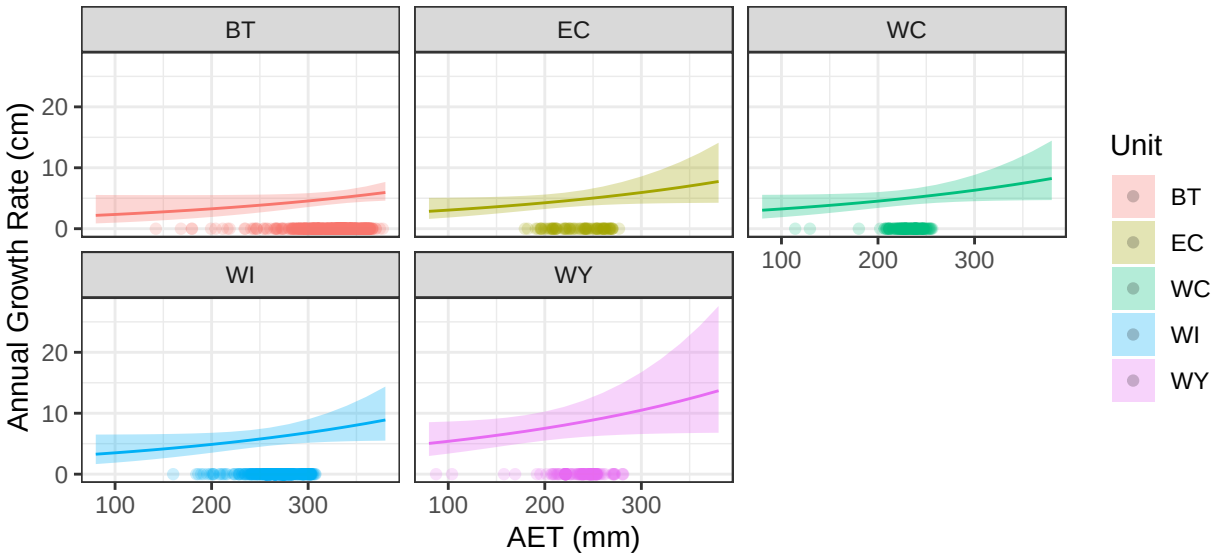


Figure 6: Predicted impact of AET of WBP seedling growth rate by planting unit

evidence that there is site-to-site variability in seedling log growth rates after accounting for the impact of AET, CWD, competition number, and unit ($\chi^2_1 = 39.925$, p = 2.638e-10). The estimated variation between sampled sites is $\sigma^2_{site} = 0.01039$ (95% CI 3.096e-03 to 1.715e-02).

4 Scope of Inference

Because there was random sampling of the WPB within each site, the relationship between AET/CWD and WBP seedling growth rate can be applied to the larger WBP population at that site, but because the sites and units were not randomly selected, no further inferences can be drawn to the larger populations of planting sites or planting units. Furthermore, this was an observational study with no random assignment of treatments, so inferences about the relationships between seedling growth rates and AET, CWD, competition, and planting unit are only correlative, not causative.

These findings can be used to identify planting locations for WBP that optimize growth rates based on modeled AET and CWD values from available climate data. Planting objectives should be considered on a unit-to-unit or site-to-site basis as there is variation in growth rates between sites and between unit.

5 Critique

Trying to follow and reproduce what Laufenberg et al. 2020 did in the paper proved to be extremely complicated. There appeared to have been extensive data cleaning and manipulation, which was not communicated in the paper. We know that they chose the "more ecologically relevant" variable when looking at correlation, but there was no clear instructions on what that was. Furthermore, the paper had no clear model selection processes, making it difficult to reproduce. Many of their choices, such as using a cubic form for AET and comp_number, were not explained and the reasoning behind was unclear. The objectives of Laufenberg et al. 2020 are unclear, and seem to combine directed questions about the role of climate and seedling environment on seedling growth rate with fishing for an explanation. Furthermore, the models selected in the paper severely violate many of the assumptions of linear models, specifically normality. Unfortunately, our simplified model also violated the normality assumption, suggesting that perhaps the dataset is missing a covariate that may help explain WBP seedling growth rate.

After trying and failing to reproduce the results from the paper, we streamlined our research question and attempted to just use the data collected to answer what we were interested in. However, even that proved to be challenging.

6 References

- Douglas Bates, Martin Maechler, Ben Bolker, Steve Walker (2015). Fitting Linear Mixed-Effects Models Using lme4. Journal of Statistical Software, 67(1), 1-48. doi: 10.18637/jss.v067.i01.
- H. Wickham. ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag New York, 2016.
- Laufenberg, David, et al. "Biophysical gradients and performance of whitebark pine plantings in the Greater Yellowstone Ecosystem." Forests 11.1 (2020): 119.

- R Core Team (2021). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.
- Thornton, P.E.; Thornton, M.M.; Mayer, B.W.; Wei, Y.; Devarakonda, R.; Vose, R.S.; Cook, R.B. Daymet: Daily Surface Weather Data on a 1-km Grid for North America; Version 3; ORNL DAAC: Oak Ridge, TN, USA, 2016.

7 Appendix

7.1 Full list of variables analyzed in study

- Age - Years since planting
- T_{mean} - Mean annual temperature (°C)
- T_{max} - Max monthly temperature (°C)
- PPT (mm) - Mean annual precipitation.
- Snowpack (mm) - Mean spring (March-May) snowpack
- WD_{annual_mean} (mm)
- WD_{annual_max} (mm)
- WD_{month_max} (mm)
- AET - Mean actual evapotranspiration (mm) during growing season (April–October)
- PET - Mean potential evapotranspiration (mm) during growing season (April–October)
- GDD - Mean annual growing degree days (April-October)
- Micro - This was a binary variable indicating presence of favorable microsite conditions. 1 if there was a rock or other topographical feature that changed the environmental conditions where the seedling lived.
- $Micro_{prop}$ - Proportion of WBP with a microsite at the site-level
- Comp_number - number of competitors within a 3.59m radius
- PICO - Presence of Lodgepole pine (*Pinus contorta*) within 3.59m radius of WBP. 1 if

PICO is present, 0 otherwise.

- PIEN - Presence of Apache pine (*Pinus engelmannii*) within 3.59m radius of WBP. 1 if PIEN is present, 0 otherwise.
- ABLA - Presence of subalpine fir (*Abies lasiocarpa*) within 3.59m radius of WBP. 1 if ABLA is present, 0 otherwise.

7.2 Figures

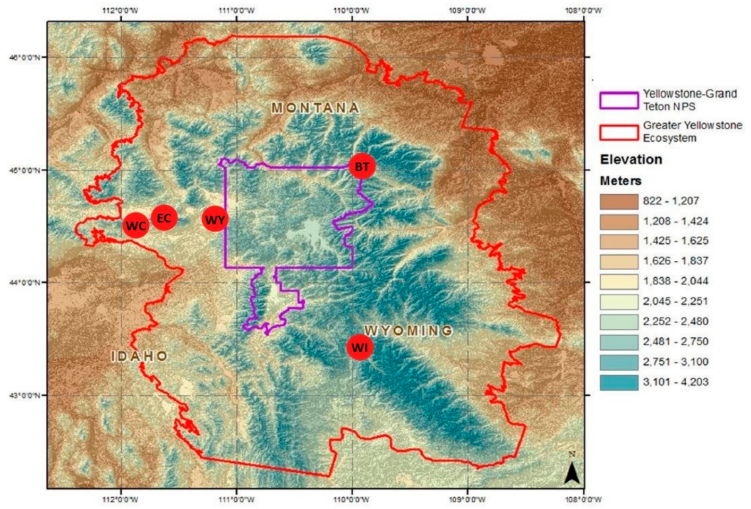


Figure 7: Map of planting units included in study in the Greater Yellowstone Ecosystem. Yellowstone National Park is outlined in purple, the greater yelowstone ecosystem is outlined in red, and the five whitebark pine planting units run by the Forest Service and National Park service are circled in red. The two-letter abbreviation denotes the name: BT = Beartooth, EC = East Centennial, WC = West Centennial, WI =Wind River, WY =West Yellowstone

7.3 Code used in Analysis

```
# this is the code script with our final data cleaning and
# analysis Steve Huysman and Parker Levinson Stat 512 April
# 14, 2023
```

```r
# load required packages #### library(raster)
library(tidyverse)
library(lme4)
library(effects)
# library(sf)
library(ggplot2)
library(kableExtra)
# library(stars)
library(lmerTest)
library(MuMIn)
library(GGally)
library(ggpubr)
library(ggcorrplot)
library(predictmeans)
library(lattice)
library(cowplot)


# read the dataframe ####
df_raw <- read.csv("./laufenberg-df.csv")


# clean it up ####
df <- df_raw %>%
    group_by(unit) %>%
    # we're not entirely sure why we have to do all these
    # adjustements, probably because the growing season is
    # 7 months long
mutate(aet = grow_aetmean * 7, cwd = grow_dmean * 7, pet = grow_petmean *
```

```
    7, p = annual_p * 12) %>%
    ungroup() %>%
    mutate(PICO = as.factor(PICO), micro = as.factor(micro),
        ABLA = as.factor(ABLA), PIEN = as.factor(PIEN))



### From Katie's comments, the above graph reproducing fig4
### is not the right raw data plot to include there is no
### reason to visualize by planting unit.  Since we are
### including site in the model as a fixed effect, I think
### we should include of a plot of aet x growth rate,
### visualized by planting site.  We can do this either by
### faceting on site or col = site in the aes(), both are
### visually noisy but the facet_wrap seems easier to read.
### from this plot, I can't see any reason to log transform
### growth rate.  title <- ggplot() + labs(title = 'AET
### (mm) vs Growth Rate (cm/year) for all study sites by
### planting unit') + theme_bw()
gridded <- df %>%
    group_split(unit) %>%
    map(~ggplot(., aes(x = aet, y = growth_rt, col = site)) +
        geom_point() + theme_bw() + theme(legend.text = element_text(size = 6),
        legend.title = element_text(size = 6)) + ggtitle(.$unit)) %>%
    plot_grid(plotlist = ., align = "hv", ncol = 2)
aet_x_growthrt <- plot_grid(gridded, ncol = 1, rel_heights = c(0.1,
    1))
```

```r
# look for correlations within variables of question
interested_var <- df %>%
    select(annual_tmax, annual_p, spring_snow, spring_rain, cwd,
        monthly_dmax, max_dsum, aet, pet, grow_gdd, annual_tmean,
        growth_rt)

interested_var.cor <- interested_var %>%
    cor()

corr_plot <- ggcorrplot(interested_var.cor, hc.order = TRUE,
    outline.col = "white", type = "lower", lab = TRUE, lab_size = 3,
    digits = 1, colors = c("#6D9EC1", "white", "maroon"), ggtheme = ggplot2::theme_bw)

selected_vars <- df %>%
    select(annual_tmax, annual_p, aet, pet, cwd, comp_number,
        micro, PICO, PIEN, ABLA, log_growth_rt)

pairs_plot <- ggpairs(data = selected_vars, upper = "blank")



# model selection with simplified model####
model_of_interest <- lmer(log_growth_rt ~ aet + age + cwd + comp_number +
    unit + annual_p + annual_tmax + PICO + ABLA + PIEN + micro +
    (1 | site), df)
null_interest <- lmer(growth_rt ~ aet + unit + (1 | site), df)
step(model_of_interest, direction = "backwards")
```

```
## Backward reduced random-effect table:
##
##            Eliminated npar  logLik    AIC    LRT Df Pr(>Chisq)
## <none>                  17 -580.29 1194.6
## (1 | site)          0   16 -587.23 1206.5 13.892  1  0.0001936 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Backward reduced fixed-effect table:
## Degrees of freedom method: Satterthwaite
##
##             Eliminated  Sum Sq Mean Sq NumDF   DenDF F value   Pr(>F)
## PICO                 1 0.01102 0.01102     1  984.97  0.0812 0.775771
## cwd                  2 0.03621 0.03621     1  558.90  0.2669 0.605593
## PIEN                 3 0.03687 0.03687     1 1066.97  0.2720 0.602126
## ABLA                 4 0.04461 0.04461     1 1174.20  0.3292 0.566215
## annual_p             5 0.19353 0.19353     1   30.60  1.4296 0.241006
## annual_tmax          6 0.06614 0.06614     1   25.55  0.4887 0.490808
## age                  7 0.13501 0.13501     1   13.13  0.9986 0.335715
## micro                8 0.21382 0.21382     1 1235.96  1.5819 0.208728
## aet                  0 0.94744 0.94744     1 1045.41  7.0039 0.008255 **
## comp_number          0 0.81612 0.81612     1  776.62  6.0332 0.014257 *
## unit                 0 2.00751 0.50188     4   26.35  3.7101 0.015970 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Model found:
```

```
## log_growth_rt ~ aet + comp_number + unit + (1 | site)
```

```
AICc(model_of_interest)  #AICc  1195.074
```

```
## [1] 1195.074
```

```
# no indication for CWD but must keep in model because it's
# our research question
reduced_model_of_interest <- lmer(log_growth_rt ~ aet + cwd +
    comp_number + unit + (1 | site), df)

AICc(reduced_model_of_interest)  #AICc 1148.384
```

```
## [1] 1148.384
```

```
### LRT
reduced_model_of_interest_no_site <- lm(log_growth_rt ~ aet +
    cwd + comp_number + unit, df)
anova(reduced_model_of_interest, reduced_model_of_interest_no_site)
```

```
## refitting model(s) with ML (instead of REML)
```

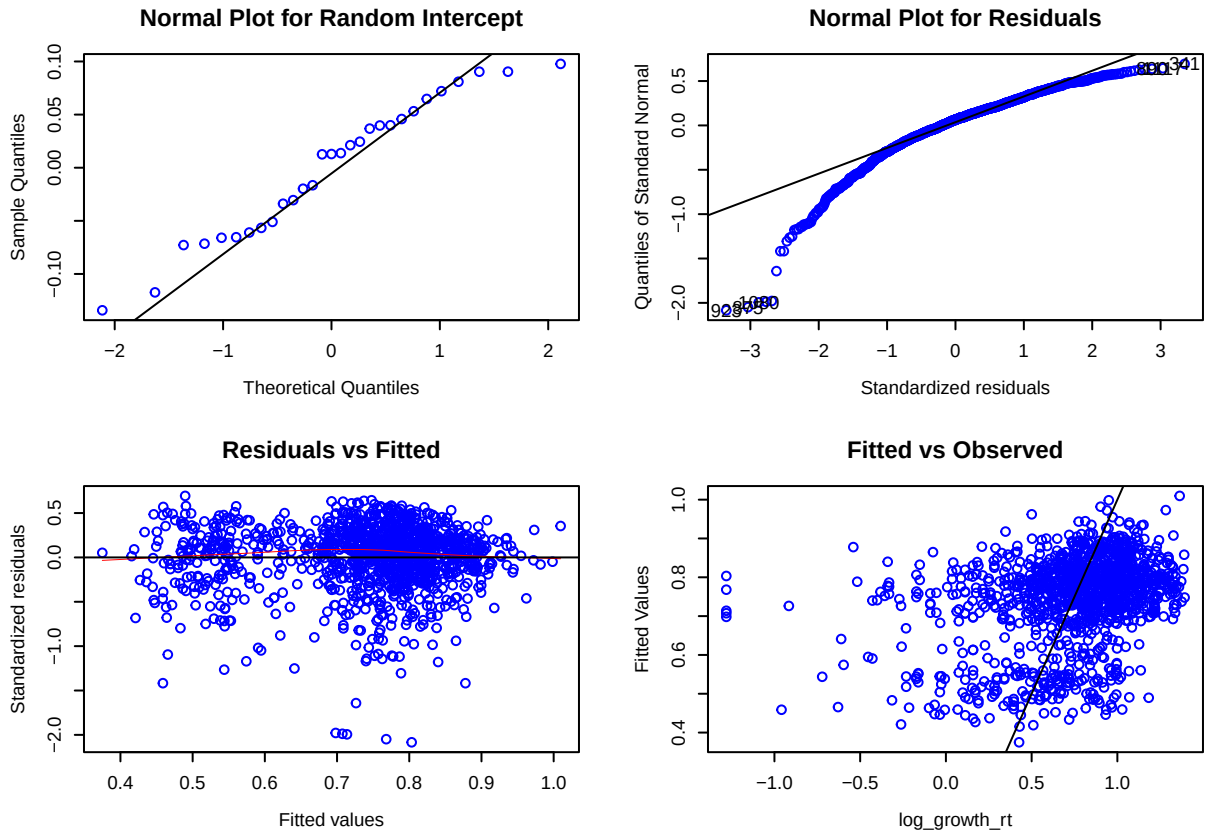
```
## Data: df
## Models:
## reduced_model_of_interest_no_site: log_growth_rt ~ aet + cwd + comp_number + unit
## reduced_model_of_interest: log_growth_rt ~ aet + cwd + comp_number + unit + (1 | site
##                                  npar    AIC    BIC  logLik deviance  Chisq Df
## reduced_model_of_interest_no_site    9 1127.4 1173.5 -554.68   1109.4
## reduced_model_of_interest           10 1089.4 1140.7 -534.72   1069.4 39.925  1
##                                  Pr(>Chisq)
```

```
## reduced_model_of_interest_no_site
## reduced_model_of_interest          2.638e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

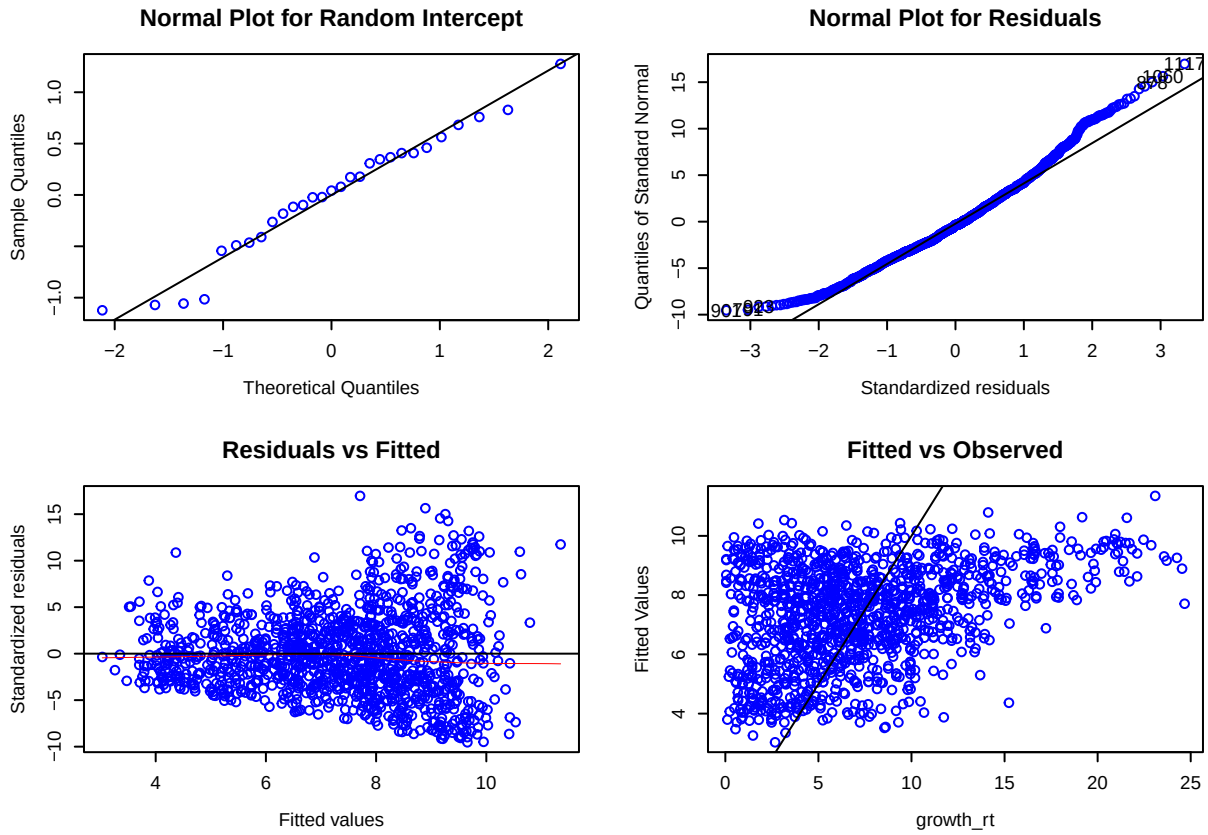
```
reduced_model_of_interest_minus_cwd <- lmer(log_growth_rt ~ aet +
    comp_number + unit + (1 | site), df)

AICc(reduced_model_of_interest_minus_cwd)
```

```
## [1] 1131.569
```

```
## 1131.569 , delta AIC = 16.815

model_no_log <- lmer(growth_rt ~ aet + age + cwd + comp_number +
    unit + annual_p + annual_tmax + PICO + ABLA + PIEN + micro +
    (1 | site), df)

# diagnostic plots####
diagnostic_log <- residplot(model_of_interest, newwd = F)
```




```
diagnostic <- residplot(model_no_log, newwd = F)
```



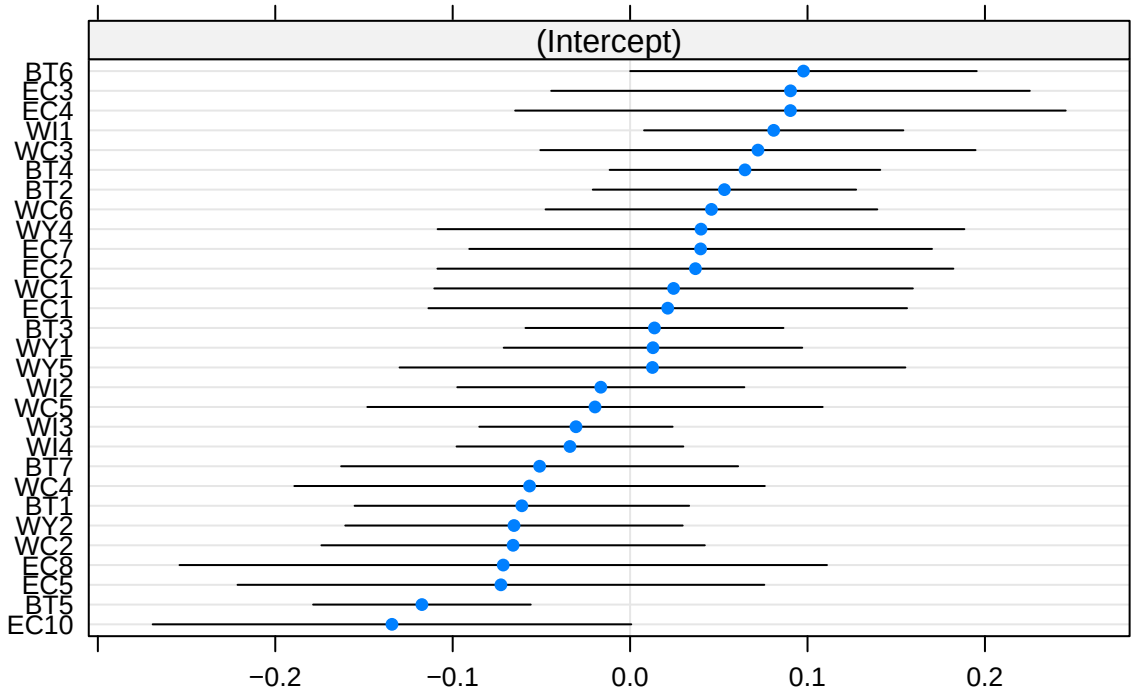

```
diagnostic_combined <- ggarrange(diagnostic, diagnostic_log)
```

```
## Warning in as_grob.default(plot): Cannot convert object of class list into a
## grob.
```

```
## Warning in as_grob.default(plot): Cannot convert object of class list into a
## grob.
```

```
# CookD(model_of_interest, newwd=T) # this won't run on the
# model this is a slow function so commented out for now,
# point 1231, 1157, and 848 are major outliers


dotplot(ranef(model_of_interest, condVar = T))  #looks relatively linear here, i don't
```

```
## $site
```




```
# create effects plot###
model_effect <- ggpredict(reduced_model_of_interest, terms = c("aet",
    "unit")) %>%
    as_tibble() %>%
    mutate(predicted_back = 10^(predicted), conf.low.back = 10^(conf.low),
        conf.high.back = 10^(conf.high), unit = group)



effect_plot <- ggplot(model_effect) + geom_line(aes(x, predicted_back,
    color = unit)) + geom_ribbon(colour = NA, alpha = 0.3, aes(x,
    predicted_back, ymin = conf.low.back, ymax = conf.high.back,
    color = unit, fill = group)) + geom_point(data = df, aes(y = 0,
    x = aet, fill = unit, color = unit), alpha = 0.2) + facet_wrap(vars(unit)) +
    labs(y = "Annual Growth Rate (cm)", x = "AET (mm)") + guides(fill = guide_legend(tit
```

```
    color = FALSE) + theme_bw()
```