Dataset Title

Greenhouse early growth data (germination to ~6 mo old), *Pinus strobiformis,* grown in Flagstaff, AZ, collected annually for different seedlots from 2015-2017

Abstract

Taken from this dissertation abstract: Bucholz, EB. (2020). Early Growth, Water Relations and Survival: Common Garden Studies of *Pinus strobiformis* under Climate Change. Doctoral Dissertation, Northern Arizona University. 166 pages.

Range-wide studies of phenotypic variation have not occurred for *Pinus strobiformis* (PIST). We used populations of PIST that represent the entire geographic range (Colorado to Southern Mexico) of PIST. Most studies to date include only a certain subset of populations from a portion of the range. We sowed seedlings in 2015, 2016 and 2017 and randomly selected a subset of 5 individuals per family to measure cotyledon length, primary needle length and cotyledon number on selected individuals 6 months after seed germination while seedlings grew in the greenhouose and assessed the importance of annual source climate variables on phenotypic expression on populations from across the entire range of PIST. Overall, we found a significant difference in growth patterns related to mean annual temperature, precipitation as snow, continentality, summer/spring precipitation balance and degree days below 0°C. These trends indicate that in Mexico, higher amounts of growing-season precipitation, longer growing season, and warmer temperatures drives increased allocation to seeds (as measured by seed weight) and increased early growth (measured by cotyledon and primary needle length) compared to US portions of the range. We found evidence for the clustering of populations on the basis of hybridization with *Pinus flexilis* in the northern part of the PIST range.

Creators

(These are the people who will show up as authors in the dataset citation. These are the individuals who have provided intellectual or other significant contributions to the creation of this dataset, much like the authors of a research paper.)

First Name	Middle Initial	Last Name	Organization	e-mail address	ORCID ID (optional)
Ethan	R	Bucholz	Northern Arizona University; Colorado State Forest Service (current affiliation)	ethanbucholz@gmail.com	
Ehren		Moler	Northern Arizona University; DroneSeed (current affiliation)	Erm287@nau.edu	
Jessica	D	Hartsell	Northern Arizona University	Jad495@nau.edu	
Kristen	M	Waring	Northern Arizona University	Kristen.waring@nau.edu	0000-0001-9935-9432
Amy	V	Whipple	Northern Arizona University	Amy.whipple@nau.edu	
Anna		Schoettle	Rocky Mountain Research Station, USDA Forest Service	Anna.schoettle@usda.gov	

Other personnel names and roles

(Who should a data user contact with questions about these data? You **must** enter a person or organization name to serve as the **contact** for this dataset. You may also list other personnel who participated in the project (such as field crew, lab tech, data entry etc.) in this table with optional fields e-mail addresses, organization and ORCID ID.)

First Name	Middle Initial	Last Name	Organization	e-mail address	ORCID ID (optional)	Role in project
Ethan	R	Bucholz	Northern Arizona	ethanbucholz@gmail.com		Contact

			University; Colorado State Forest Service (current affiliation)			
Kristen	M	Waring	Northern Arizona University	Kristen.waring@nau.edu	0000-0001-9935-9432	Lead PI
Anna		Schoettle	Rocky Mountain Research Station, USDA Forest Service	Anna.schoettle@usda.gov		*P. flexilis* data contact / creator

License

(Select a license for release of your data. We have 2 recommendations: CCO – most accommodating of data reuse, & CCBY – requires attribution).

CCBY

Keywords

(**List keywords below and separate with commas.** Using keywords from a controlled vocabulary (CV) will improve the future discovery and reuse of your data. The LTER CV is a good source for keywords. **Access the LTER CV here**. Also, please determine one or two keywords that best describe your lab, station, and/or project (e.g., Trout Lake Station, NTL LTER).)

Plant properties (germination, seedling traits and growth), climate-growth relationships, *Pinus strobiformis*, *Pinus flexilis*, climate change, greenhouse, NAU Department of Biological Sciences, NAU School of Forestry, NAU Silviculture and Applied Forest Health Lab, USDA Forest Service Rocky Mountain Research Station

Funding of this work:

List only the **main PI of a grant** that supported this project, starting with the main grant first. Add rows to the table if several grants were involved.

PI First Name	PI Middle Initial	PI Last Name	PI ORCID ID (optional)	Title of Grant	Funding Agency	Funding Identification Number
Kristen	M	Waring		NSF Grant 1	National Science Foundation	EF-1442597

Timeframe

- Begin date: 01/2015

- End date: 08/2017
- Data collection ongoing/completed: 07/2017

Geographic location

(Use **decimal degrees** to define a point or a bounding box. Use a negative symbol (-) to indicate a west longitude. Copy this block to add multiple points or areas.)

- Verbal description: The entire range of *Pinus strobiformis* from SW Colorado to southern Mexico. Seeds collected from 3-5 maternal trees per population (progeny from a single maternal tree termed 'family' while 3-5 maternal trees comprise the population). Seed collected between 2012-2016 from maternal trees across the entire range of *Pinus strobiformis*. We also received data from Rocky Mountain Research Station research ecologists on *Pinus flexilis* seedlings with resistance to white pine blister rust that was folded into this dataset.
- North bounding coordinate: 42.0
- South bounding coordinate: 22.0
- East bounding coordinate: -102.5
- West bounding coordinate:-113.0

Taxonomic species or groups

(Does your data focus on particular taxa? If so, please list them here.)

Pinus strobiformis Engelm.

Pinus flexilis E. James

Methods

(Be specific about the study design and field and lab methods for collecting and processing the data. Include instrument descriptions and protocol citations.)

From Ethan Bucholz Dissertation Chapter 2 (see top for citation)

Seed Sources and Study Implementation

Between 2012-2016 we collected PIST seed from Colorado, Arizona, New Mexico, Texas and Mexico (Table 2.1). Selected populations spanned 16.9 degrees latitude and 7.7 degrees of longitude. Seed collection locations included 10 US National Forests, 1 US National Park, 2 Native American Reservations, 1 Nature Conservancy Preserve, 32 ejidos (Mexican community forests) and 8 privately owned forests in Mexico. Populations consisted of 3-5 maternal trees. Across three years (2015, 2016, 2017), families representing the entire PIST range (66 populations, 291 families) were sown in a greenhouse for later out-planting in a related common garden study (Table 2.1, Figure 2.1). After collection and processing of seed cones, average dry seed weights were determined for each family using an average of 10-seed lots in 2015 and 2017. In 2016, individuals seeds were weighed prior to sowing. We also used data recorded on 14 populations of PIFL grown in a greenhouse (Figure 2.1). Populations and families selected were considered either PIST or PIFL *a priori*. Given the recent indications that PIST and PIFL readily hybridize (Menon et al. 2018) geographic overlap in PIST and PIFL can be seen in southwestern Colorado (Figure 2.1).

Prior to sowing, seeds were surface sterilized for 24 hours in a 1% hydrogen peroxide solution followed by cold stratification for 4 weeks at 4°C. In mid-January 2015, 2016 and 2017, seeds representing the entire PIST range were sown in the Northern Arizona University experimental greenhouse over the course of 1 week. Forty-cell Tinus book planters were randomly allocated to one of 3 common gardens, and seeds were then randomized and sown within each planter. Depending on expected germination (from seed x-rays and previous knowledge of seed lots), seeds were either single, double or triple sowed in each sow year. Greenhouse temperatures were maintained at a night-time temperature of 15 °C and at 26°C during the day, and a constant relative humidity of 50%. Seedlings were watered three times weekly to soil capacity. At the start of the growing season, plants were fertilized twice a week with a 20-20-20 N-P-K solution starting at 15ppm. Each week following, the solution was increased by 15ppm until it reached 60ppm (4 weeks).

Limber pine seed was collected in 2012 from each of 13 populations spanning 36.7°N to 41.4°N (see Borgman et al. 2015). An additional southern site was included with seed collected from three families in 2003, 2005, and 2006. Each population was represented by seeds from each of three open-pollinated PIFL mother trees, spaced at least 60 m apart (see Borgman et al 2015). After collection and processing of seed cones, average seed weights were determined for each family using an average of 10-seed lots.

Limber pine seed was stratified for 6 weeks at 1-2 ^{o}C, and then sown in an experimental greenhouse in Colorado on moistened filter paper (18 °C in the light and 16 °C in the dark, 12-h photoperiod; Precision Low Temperature Illuminated Incubator 818, Thermo Fisher Scientific Inc., Waltham, Massachusetts) in March 2013. Once germinated (2 mm radical emergence), 14 germinants per mother tree were immediately transplanted into 656 mL Deepots D40h (Stuewe and Sons, Inc., Tangent, Oregon) in a mixture of 20% forest soil, 50% Fafard 4P mix potting soil (Conrad Fafard Inc., Agawam, Massachusetts), 20% sand, and 10% pea gravel for drainage (see Borgman et al. 2015). Each pot was top dressed with Osmocote Classic 14–14–14 controlled-release fertilizer (Everris International B.V., the Netherlands). Greenhouse temperatures varied between 17 °C and 22 °C, with supplemental lighting providing a 16 h light – 8 h dark photoperiod. Seedlings were well-watered receiving 50 mL of deionized water per pot each week.

Measurements

In June and July 2015, 2016 and 2017, measurements on PIST were recorded once on a randomly selected subset of five seedlings/family/assigned common garden (n=15/family measured in the greenhouse). We measured cotyledon length (mm), number of cotyledons (count), and the length of three randomly selected primary needles (mm). All primary needle and cotyledon lengths were measured with digital calipers (L.S. Starret Company, MA, USA). Dead seedlings were replaced with randomly selected living seedlings to maintain the sample size of 15 per family. In 2016, we assessed time to germination after initial sowing, as each cell in this year was sown with a single seed. We did not assess time to germination in 2015 and 2017 due to sowing multiple seeds per cell .

Cotyledon and primary needle length of the limber pine seedlings were measured 116 days post-germination (July 2013) with a digital caliper on each of the 14 seedlings in each family and averaged (see Borgman et al 2014).

Climate Variables

To compare source climate relationships with phenotypic traits of interest, we used the Parameter-elevation Regression on Independent Slopes Model (PRISM) to calculate climate normals from the period of 1981-2010 (Oregon State University). Climate NA (v. 5.21) uses historical data from the dataset CRU-TS3.22 to interpolate historical climatic information for a given set of geographic points at 1km resolution (Wang et al. 2016). We extracted annual climate information from Climate NA for each family (Table 2). These climate variables, and additional derived parameters are the same as those used by Shirk et al. (2018) (Table 2.2).

Data Analysis

Objective 1: Measured Traits and Source Climate Relationships

Multicollinearity, or redundant variables, create problems with classification accuracy in feature-selection analysis like Random Forests (e.g. Kubus 2019). Feature elimination is therefore proposed to reduce errors in classification accuracy that arise from the inclusion of redundant variables in feature-selection analysis (e.g. Gregorutti et al. 2017). To reduce multicollinearity among the source climate variables, we created a list of the most related variables (Pearson's correlation coefficients over 0.90 or below -0.90). We then reduced the climate variables present in our models *apriori* based on literature support for certain variables as well as eliminating variables based on correlation coefficients >|0.9|. (e.g. Griesbauer et al. 2011; McKown et al. 2014; Goodrich et al. 2016; Kapeller et al. 2016). We used this reduced climate variable dataset (Table 2.2) for all following analyses. We used Principle Components Analysis to qualitatively describe relationships between the environmental variables and our measured variables. All measured trait values and environmental parameters were scaled and averaged by population before ordination analysis. Ordination of the data was conducted with PCOrd version 6 (McCune and Mefford 2011).

The random forest algorithm (Breiman 2001) was used to construct bootstrap samples of 9,999 "forests" from our original data set. These regression "trees" form nodes wherein a random sample (mtry) of the predictors at each node determine the best split of the predictor variables (Liaw and Wiener 2002). We then calculated scaled variable importance scores, which are derived from the nodes of "trees" created by the algorithm, wherein an important variable does not increase or marginally increases mean error of the tree (Genuer et al. 2010). Our models used in random forests contained all climate, topographic and edaphic features as predictors, with seed weight, cotyledon length, average primary needle length and cotyledon number as response variables, to determine which predictors were the most important in driving morphological variation. Higher values of scaled variable importance indicate those climate features most strongly associated with the response variables (e.g.Bhuyan et al. 2017; Li et al. 2017; Breiman 2001). We used the 'boruta' algorithm to assess variable performance by comparing relevance of predictors to data-driven randomly developed predictors (Kursa and Rudnicki 2010). Boruta compared performance of actual predictors in the dataset to randomly computed predictors to assess how each predictor performs relative to the distribution of the minimum, mean and maximum values from randomly permuted observations. Variable importance is compared to these z-score values to (alpha=0.01), to determine those attributes confirmed important. Importance values for predictors are reported. From our random forest model, we built partial dependency plots to assess how particular variables impacted trait responses. Partial dependency plots represent marginal effects of chosen predictors on the response variable, while accounting for the other variables present in the model (Friedman and Popescu 2008). Partial dependency plots graphically represent the impacts of predictors on response variables, while maintaining the effects of other predictors from the model. They also demonstrated the type of relationship (e.g. linear, non-linear etc).

Early growth data for PIFL were analyzed separately using the random forests algorithm due to differences in sample sizes between populations of PIST and PIFL, as well as the different sampling methodology employed to measure early growth characteristics of PIFL. Being that cotyledon number was not counted for PIFL, these data were excluded from our analysis. PIFL data used in this study were part of a previous study (Borgman et al. 2014) of maternal climate effects on growth, and therefore a different methodology which did not include cotyledon number. All model fitting and testing were conducted within the R statistical programming environment (v. 3.4.2 R Core Development Team 2017).

Objective 2: Introgression and Population Clustering

To assess whether seedling phenotypic variation characterizes hybridization and introgression between PIST and PIFL, we used hierarchical cluster analysis of scaled environmental variables and measured phenotypic trait variation to identify clusters within the species. With a Euclidian distance matrix and Ward's D2 method, we assessed the number of major clusters in our dataset to find support for environmental variation between cluster sets that may drive morphological differences. We then compared our geographical group cluster split to that of the core and periphery PIST populations, as reported by Menon et al. (2018), to see if there was a pattern of trait variation related to the hybridization. Using a paired t-test (alpha=0.05) we compared cluster group trait means to assess significant differences among groups.

Data Provenance

(Were these data derived from other data? If so, you will want to document this information so users know where these data came from. Please specify the source datasets used in the below **provenance table**, preferably with their DOI or URL. An example of a dataset derived from several others is here.)

Dataset title	Dataset DOI or URL	Creator (name & email)	Contact (name & email)

Data Table

(Provide a Table Name and Table Description. Each row in the below table describes one column in your data table. Complete each row as follows:

- **Column name**: This name must be exactly as it appears in the dataset. Please avoid special characters (like & or \), dashes and spaces. Underscores are permissible. Do not begin a column name with a number.
- **Description**: Please give a specific definition of the column name. This can be lengthy.
- **Unit:** Identify units for all numeric variables. Please avoid special characters and describe units in this pattern: e.g. microSiemenPerCentimeter, microgramsPerLiter, absorptionPerMolePerCentimeter

- **Code explanation**: If you use codes in your column, please explain in this way: e.g., LR=Little Rock Lake, A=Sample suspect, J=Nonstandard routine followed
- **Date format**: Please tell us exactly how the date and time is formatted: e.g. mm/dd/yyyy hh:mm:ss plus the time zone and whether or not daylight savings was observed. ISO date format of YYYY-MM-DD or YYYY-MM-DD hh:mm:ss is preferred.
- **Missing value code**: If a code for 'no data' is used, please specify: e.g., -99999

Table name: Early growth x population annual climate variables
Table description: This data table is the main data table for all populations sampled and averaged for measured characteristics and population climate variables. It is not the raw data but a compiled average based on the raw data.

Column name	Description	Unit or code explanation or date format	Missing value code
Site_Code	2,3 or 4 letter abbreviation of population name	See description	NA
Cot_Num	Population average number of cotyledons counted	Number/see description	NA
Cot_Len	Population average cotyledon length	mm	NA
Prim_Need1	Average primary needle 1 length	Mm	NA
Prim_Need2	Average primary needle 2 length	Mm	NA
Prim_Need3	Average primary needle 3 length	Mm	NA
Avg_Prim	Population average primary needle length total	mm	NA
Seed_Weight	Average mass of seeds (g)	Grams	NA
Latitude	Latitude of population (decimal degrees)	Decimal degrees	NA
Longitude	Longitudee of population (decimal degrees)	Decimal degrees	NA
MAT	Mean annual temperature (°C)	See description	NA
MWMT	annual Mean warmest month temperature (°C)	See description	NA
TD	annual Continentality; temperature difference between MWMT and MCMT (°C)	See description	NA
MAP	Mean annual precipitation (mm)	See description	NA
AHM	annual heat-moisture index (MAT+10)/(MAP/1000))	See description	NA
SHM	summer heat-moisture index ((MWMT)/(MSP/1000))	See description	NA
DD_0	annual Degree days below 0°C	See description	NA
bFFP	annual the day of the year on which FFP begins	See description	NA
FFP	annual frost-free period	See description	NA

PAS	annual precipitation as snow (mm)	See description	NA
EMT	extreme minimum temperature over 30 years (°C)	See description	NA
EXT	extreme maximum temperature over 30 years (°C)	See description	NA
CMD	annual Hargreaves climatic moisture deficit (mm)	See description	NA
MAR	mean annual solar radiation (MJ m-2 d-1)	See description	NA
RH	mean annual relative humidity (%)	See description	NA
SMRSPB	summer/spring precipitation balance (summer PPT/spring PPT)	See description/index/no units	NA

(Copy this table to document more than one data table.)

Spatial data objects

(List any geospatial data objects you would like to archive. Organize spatial data into .zip directories and describe each.)

Directory name: (A short name for the data)
Directory description: (A brief description of the data)

Attribute	Value
Horizontal Coordinate System Name (e.g. WGS_1984_UTM_Zone_12N)	
Horizontal Accuracy Report	
Vertical Accuracy Report	
Cell Size X Direction	
Cell Size Y Direction	
Raster Origin (e.g. Upper Left)	
Number of Rows	
Number of Columns	
Number of Verticals	
Cell Geometry (e.g. pixel)	

Scripts/code (software)

(List any software scripts/code you would like to archive along with your data. These may include processing scripts you wrote to create, clean, or analyze the data.)

File name	Description	Scripting language

Other objects (misc.)

(List any other objects (e.g. .zip, .pdf, etc.) you would like to archive.

File name	Description	Data type

Articles

(List articles citing this dataset)

Article DOI or URL (DOI is preferred)	Article title	Journal title

Notes and Comments