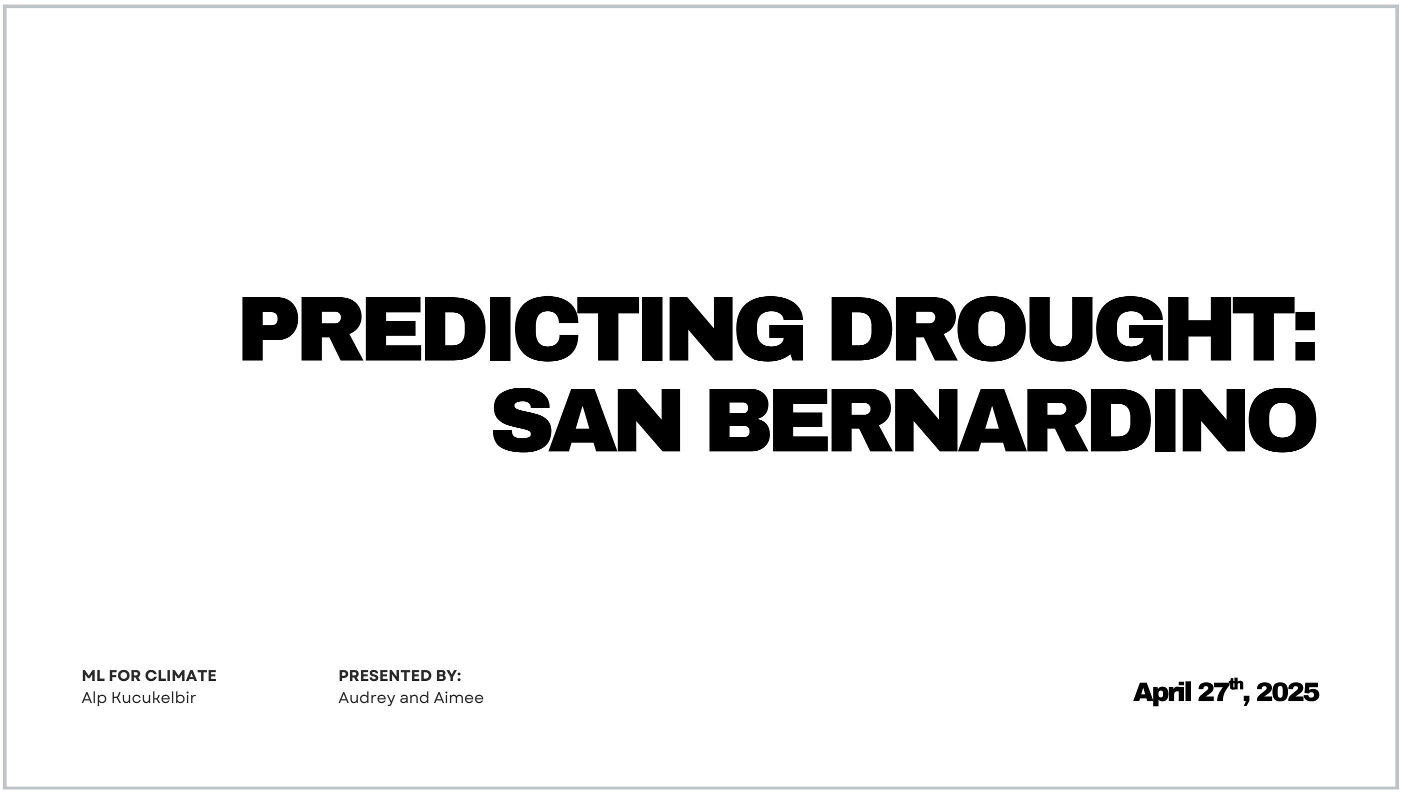


PREDICTING DROUGHT: SAN BERNARDINO

ML FOR CLIMATE
Alp Kucukelbir

PRESENTED BY:
Audrey and Aimee

April 27th, 2025

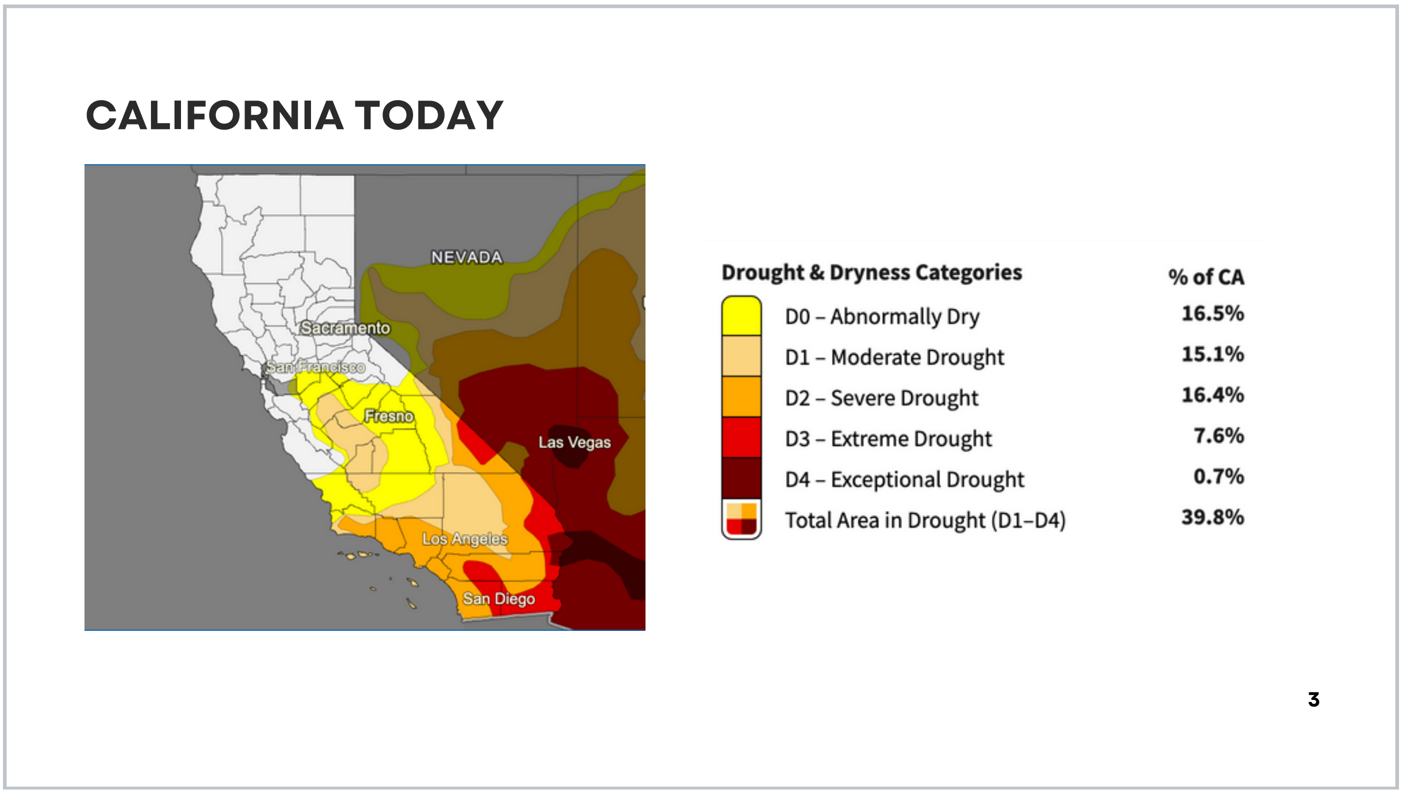
Hi, this is Aimee and Audrey, and our project is about predicting drought in the Southern Californian county of San Bernardino.

OBSERVATION & MOTIVATION

Where and why are we looking at drought?

2

So what motivated us to look at droughts, and why have we chosen California?



CALIFORNIA TODAY

	Drought & Dryness Categories	% of CA
🟨	D0 – Abnormally Dry	16.5%
🟧	D1 – Moderate Drought	15.1%
🟧	D2 – Severe Drought	16.4%
🟥	D3 – Extreme Drought	7.6%
🟫	D4 – Exceptional Drought	0.7%
▨	Total Area in Drought (D1–D4)	39.8%

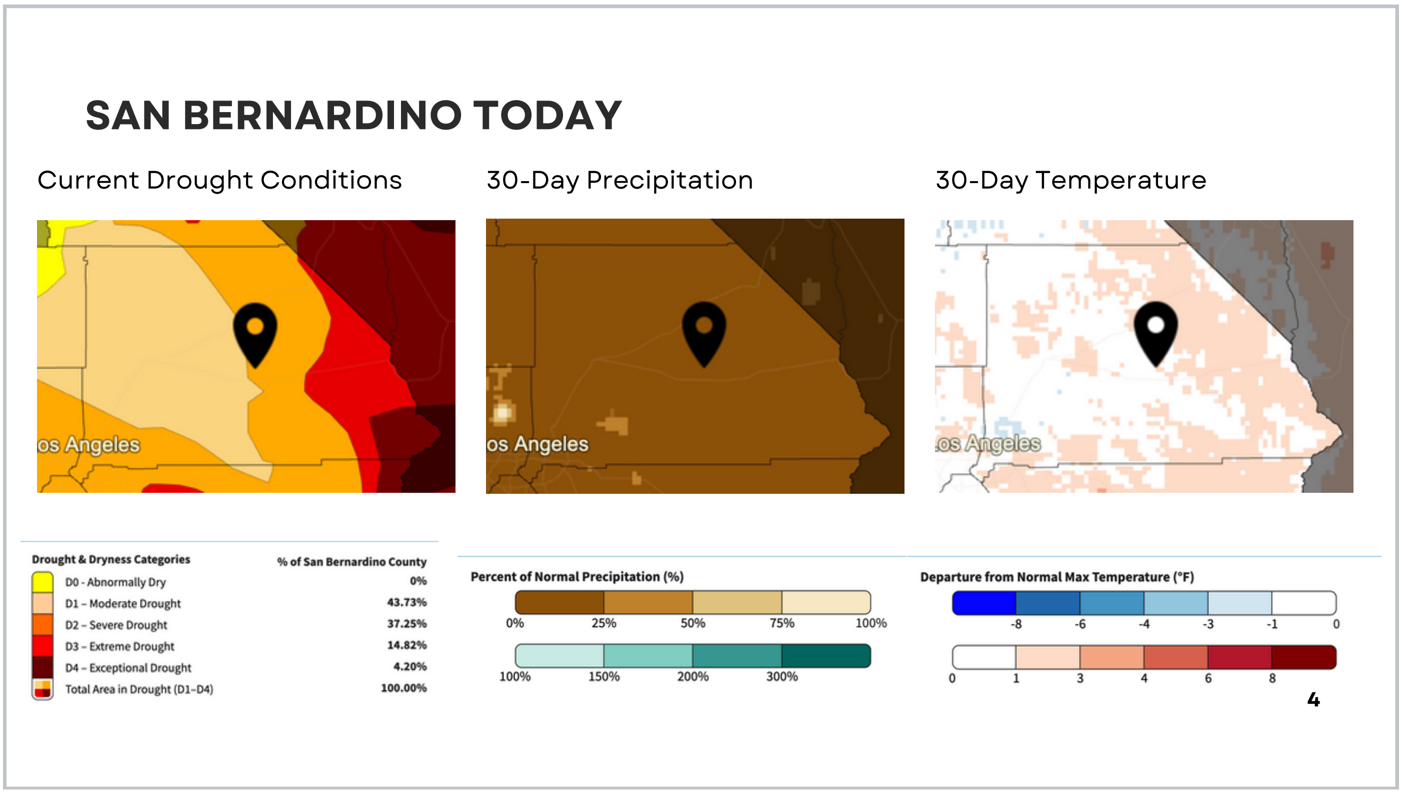
3

California, while more recently in the news for all of its wildfires, is known historically as being prone to drought, across the entire state, but especially in the south and central valley.

Drought conditions have been recorded for at least the last century, with recent events including
- 5 year statewide drought from 2012-2016
- a few from 1976-1977, 1987-1992, and 2007-2009
- dry conditions spanning more than a decade in the 1920s and 1930s

This is especially important
- given the 22.7 Million people living in a drought in the state (as of 2025)

This is really interesting because it gives us a century span of insight:
- are they recurring, cyclical? They tend to last anywhere from a year to a decade at a time, what exacerbates them?
- are the droughts worsening in severity as we look into the future?
- what is the impact of our emissions on all this (is climate the only deciding factor, since we can't control precipitation etc)



SAN BERNARDINO TODAY

Current Drought Conditions

30-Day Precipitation

30-Day Temperature

Drought & Dryness Categories	% of San Bernardino County
🟨 D0 - Abnormally Dry	0%
D1 – Moderate Drought	43.73%
D2 – Severe Drought	37.25%
🟥 D3 – Extreme Drought	14.82%
D4 – Exceptional Drought	4.20%
Total Area in Drought (D1–D4)	100.00%

Percent of Normal Precipitation (%)

Departure from Normal Max Temperature (°F)

4

Our focus: San Bernardino

- large agricultural state (https://data.nass.usda.gov/Publications/AgCensus/2022/Online_Resources/County_Profiles/California/cp06071.pdf)

- mainly crop and pasture

- large county that is currently 100% in a drought and represents 4 of the drought severities throughout (d1-d4)

- we're looking for correlations and impacts from different variables like temperature and precipitation

- despite the differences in drought severity, we can see that precipitation is fairly equal across the entire state
- it can't be the only variable that factors into the drought

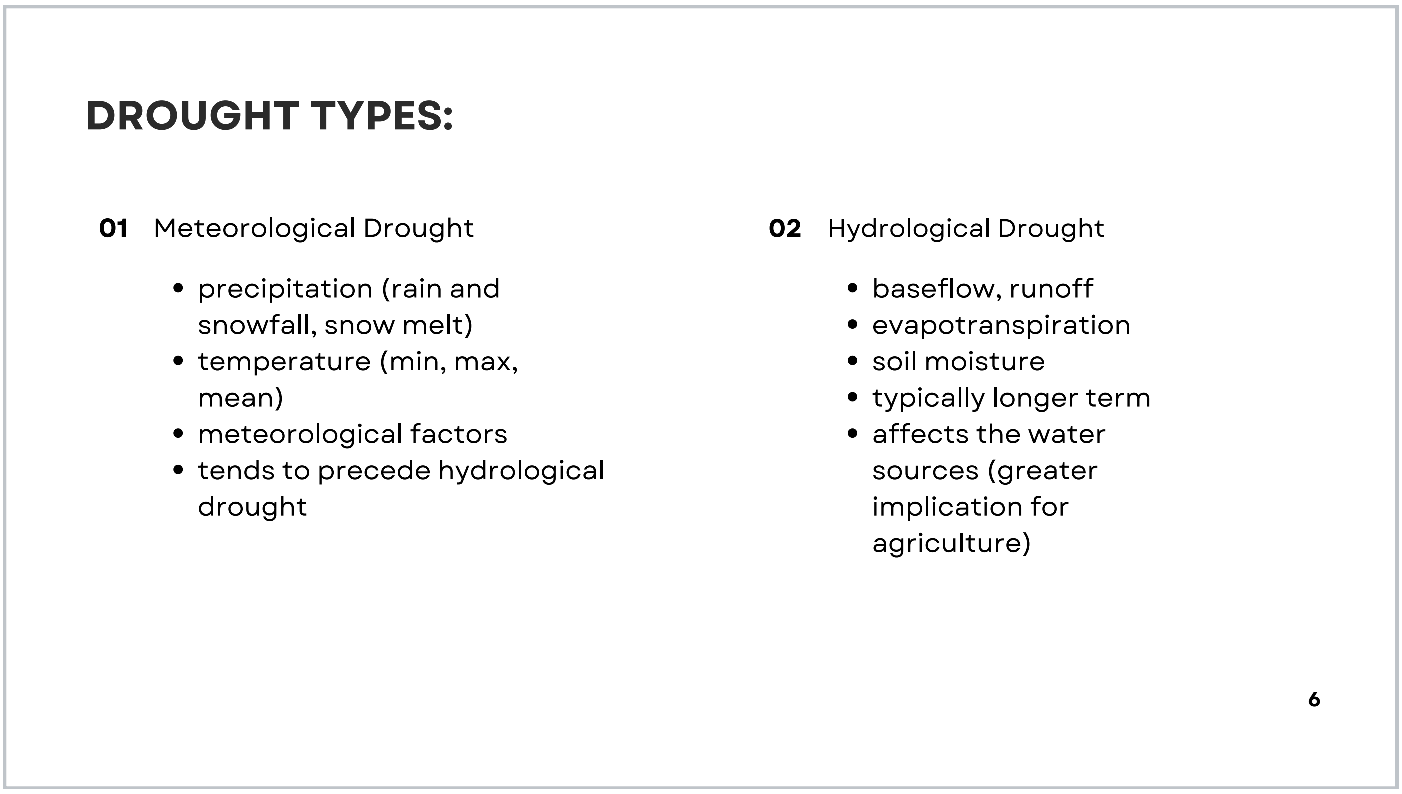
- temperatures seem to match up a bit more to the regions of d3-d4 drought severity from a few images ago
- but again, it can't be the only variable that factors in

DEFINING DROUGHT

What kinds are there?

5

In order to determine what variables to look into for our prediction model, we need to define drought more clearly and (beyond these severity labels)



DROUGHT TYPES:

01 Meteorological Drought

- precipitation (rain and snowfall, snow melt)
- temperature (min, max, mean)
- meteorological factors
- tends to precede hydrological drought

02 Hydrological Drought

- baseflow, runoff
- evapotranspiration
- soil moisture
- typically longer term
- affects the water sources (greater implication for agriculture)

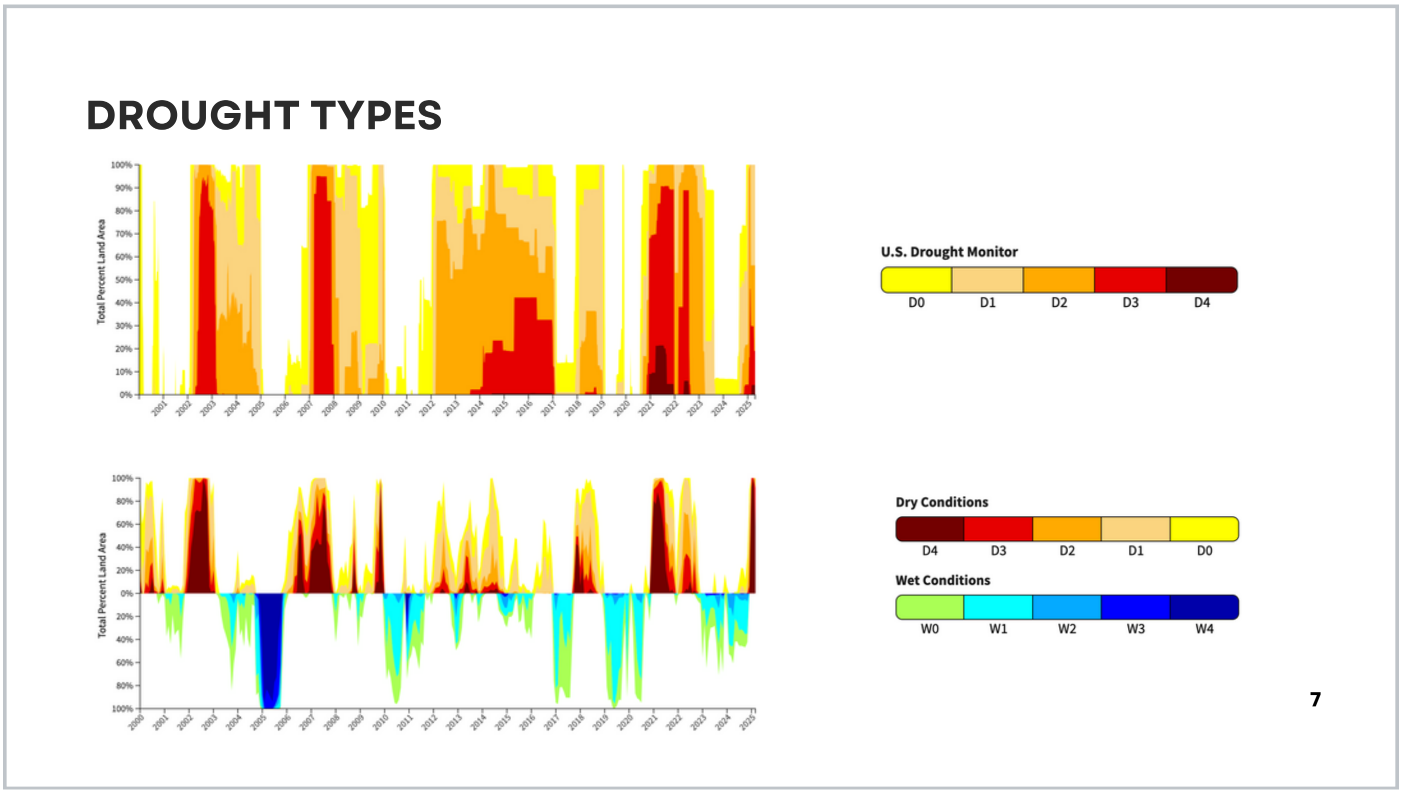
6

In its most basic sense, drought is the absence of water, which is typically quite slow in its development with an undefined beginning and end (compare it to other events such as hurricanes that are clearly visible and time constrained).

(Not to be confused with aridity, where the demand of water supply evapotranspiration, is greater than supply precipitation, which tends to be permanent in terms of climate, rather than temporary).

There's multiple scales of drought
- meteorological = abnormally dry weather with little precipitation
- hydrological = effects of meteorological drought on water availability
- agricultural = when meteorological and hydrological drought conditions affects our ability farm (soil moisture deficits, irrigation water availability, pasture impact, crop reduction)
- socioeconomic -> affects human supply and demand of commodities affected by drought
- ecological

WE'RE LOOKING AT THE FIRST TWO, precursors to other types of drought, meteorological tends to precede hydrological



DROUGHT TYPES

7

Also in terms of droughts:

there is a constant state of drought severity, but can we determine these large drought periods (and possibly their types?)

we can see the 2007-2009, 2012-2017 droughts, 2020-2022 (for all of california but here it seems more like 2021-2023)

and the 2025 drought

- are they recurring, cyclical? They tend to last anywhere from a year to a decade at a time, what exacerbates them?
- are the droughts worsening in severity as we look into the future?
- what is the impact of our emissions on all this (is climate the only deciding factor, since we can't control precipitation etc)

can we predict these drought blocks and see if they really are increasing in duration and intensity over time and across the century (see the presence of more D4 land in this decade than the 2 previous ones)

here we show drought vs the dry/wet conditions graphs, clearly the droughts are getting counterac ted/resolved (i am not a climate scientist).

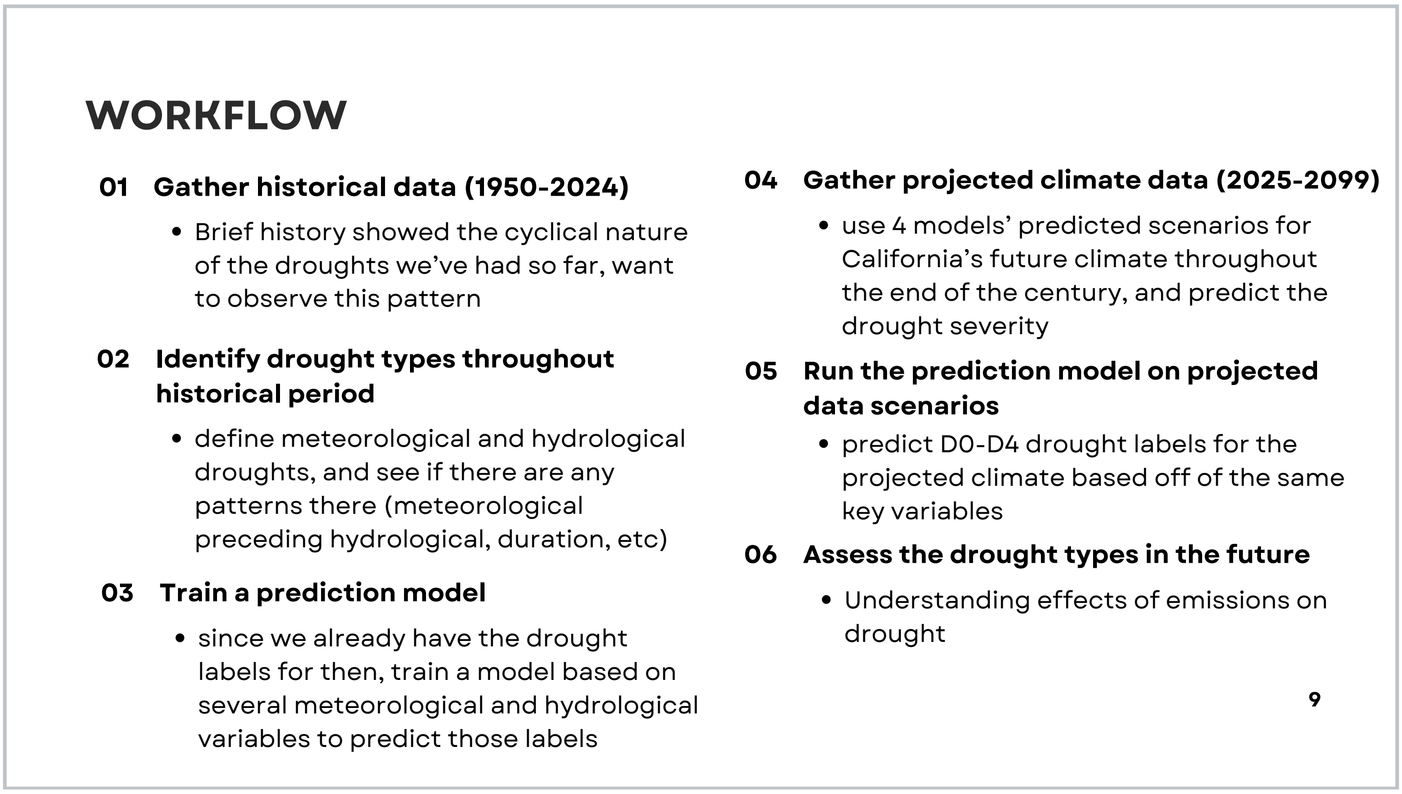
There are many types of drought (meteorological, hydrological, agricultural, socioeconomical https ://drought.unl.edu/Education/DroughtIn-depth/TypesofDrought.aspx)

1) meteorological drought, 2) hydrological drought, 3) agricultural drought, and 4) socioeconomic d rought.

Meteorological drought refers to the occurrence of drought due to insufficient precipitation. Agricultural drought primarily stems from insufficient soil moisture and then consequently lead to crop yield reduction (Mishra et al. 2010). Agricultural drought is closely related to meteorological drought (Barker et al., 2016). This is because prolonged periods of inadequate precipitation can lead to a decline in soil moisture, thereby causing the occurrence of agricultural drought.

THE GAMEPLAN

How do you go about predicting climate?



WORKFLOW

01 Gather historical data (1950-2024)

- Brief history showed the cyclical nature of the droughts we've had so far, want to observe this pattern

02 Identify drought types throughout historical period

- define meteorological and hydrological droughts, and see if there are any patterns there (meteorological preceding hydrological, duration, etc)

03 Train a prediction model

- since we already have the drought labels for then, train a model based on several meteorological and hydrological variables to predict those labels

04 Gather projected climate data (2025-2099)

- use 4 models' predicted scenarios for California's future climate throughout the end of the century, and predict the drought severity

05 Run the prediction model on projected data scenarios

- predict D0-D4 drought labels for the projected climate based off of the same key variables

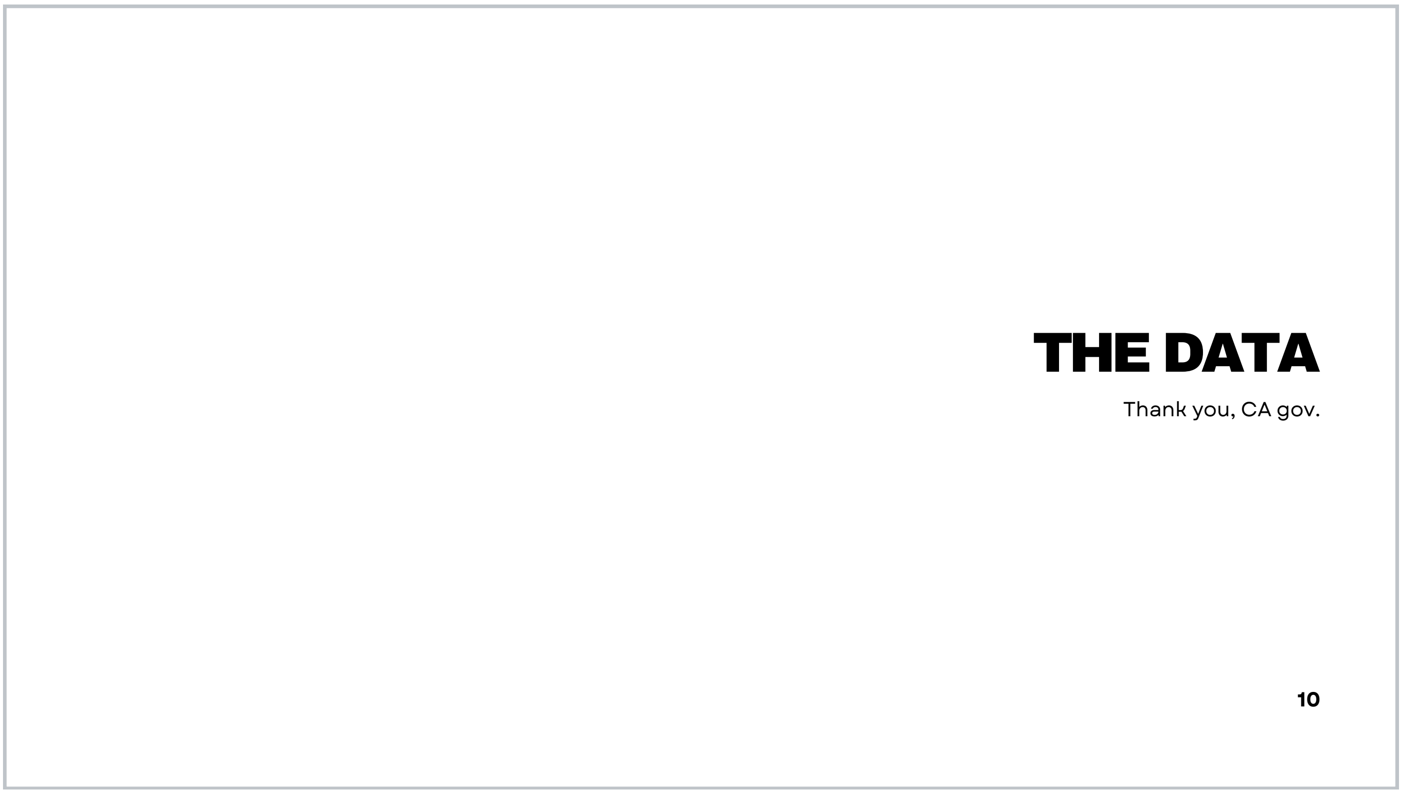
06 Assess the drought types in the future

- Understanding effects of emissions on drought

drought is a complex system of variables

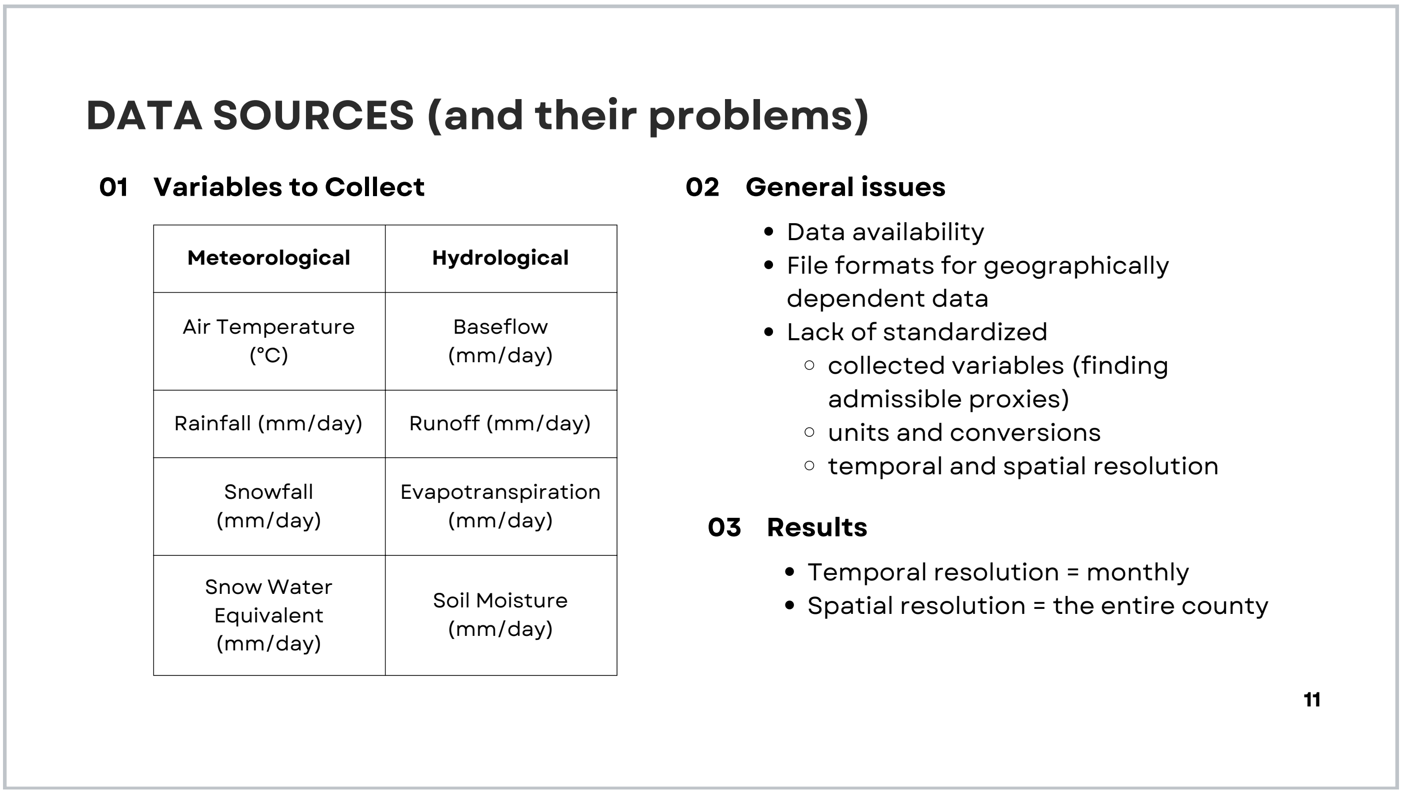
How do we plan on predicting drought level values for San Bernadino?

gather historical data
train a model using the features in the historical data



THE DATA

Thank you, CA gov.



DATA SOURCES (and their problems)

01 Variables to Collect

Meteorological	Hydrological
Air Temperature (°C)	Baseflow (mm/day)
Rainfall (mm/day)	Runoff (mm/day)
Snowfall (mm/day)	Evapotranspiration (mm/day)
Snow Water Equivalent (mm/day)	Soil Moisture (mm/day)

02 General issues

- Data availability
- File formats for geographically dependent data
- Lack of standardized
    - collected variables (finding admissible proxies)
    - units and conversions
    - temporal and spatial resolution

03 Results

- Temporal resolution = monthly
- Spatial resolution = the entire county

11

Given our drought definitions earlier, we decided to collect the data on the following 8 variables.
- READ TABLE

Pretty much all of the available data comes from either universities or government agencies, at different levels (state vs county, etc).

As we've discussed a lot this semester, there's a lot of issues with this distributed data collection, and we witnessed it firsthand:

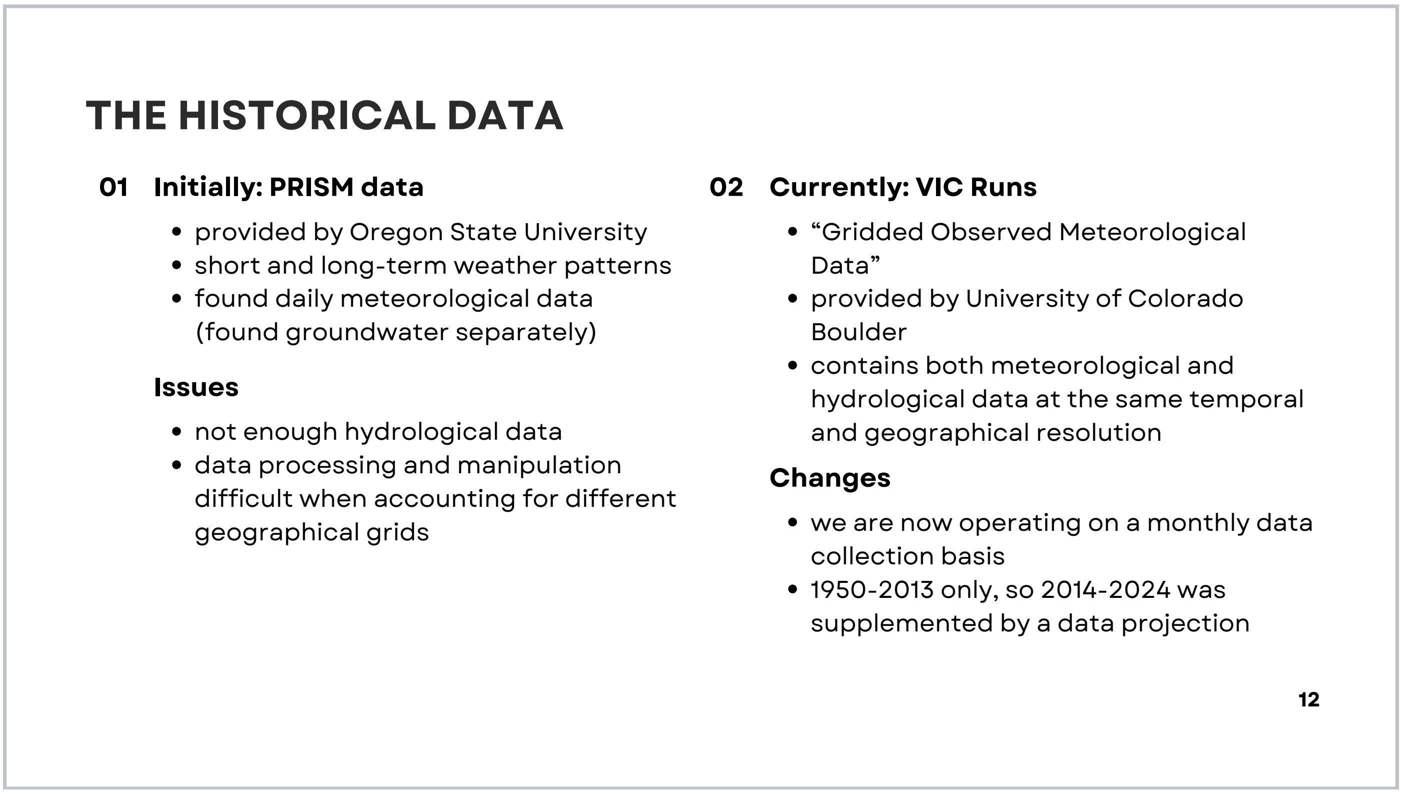
There's a lack of standard variables and units of measurement
- file structure: data is so location dependent (and better understood as heatmaps overlaid on geographic maps), that they are typically available in GEOTIFF format, from which extracting CSV files is tricky just from the size of geotiff and needing to align shapefiles for the correct county

- time periods: is this being measure daily, weekly, monthly, at what time of the day? This matters a lot for meteorological data that can vary so quickly throughout the day

- variables: different variables can act as "proxies" for one another: some will measure relative humidity, others dew point (relative humidity is the ratio of amount of water vapor in the air to the maximum amount the air can hold at a given temperature. Dew point is the temperature at which the air, at a constant atmospheric pressure, must be cooled to become saturated, meaning it can't hold any more water vapor) -> It's difficult finding the appropriate conversion rates for these proxies and have your model use them interchangeably

- resolution: what's the grid resolution of the ground covered by that variable?

- location restrictions: groundwater is only measured at selected locations and wells



THE HISTORICAL DATA

01 Initially: PRISM data

- provided by Oregon State University
- short and long-term weather patterns
- found daily meteorological data (found groundwater separately)

Issues

- not enough hydrological data
- data processing and manipulation difficult when accounting for different geographical grids

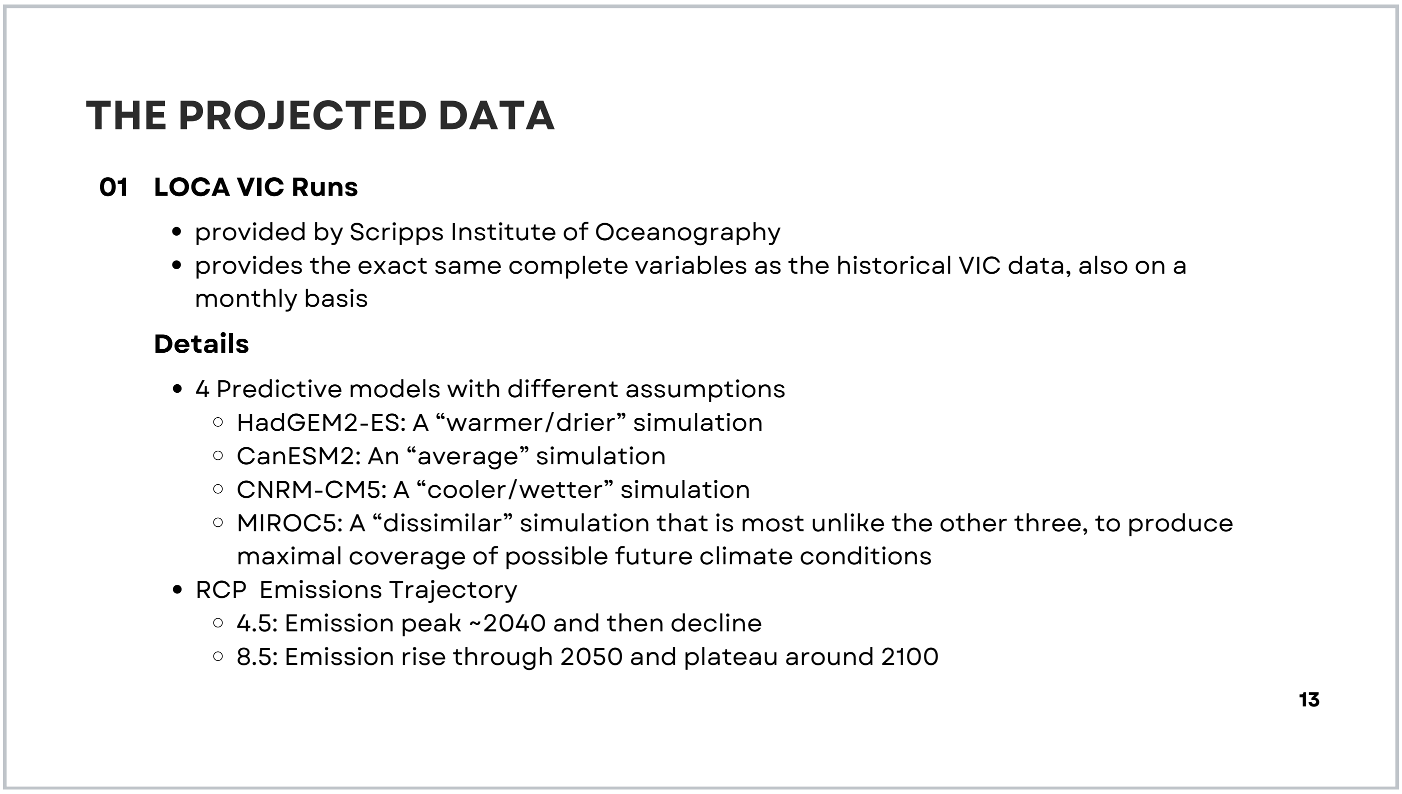
02 Currently: VIC Runs

- “Gridded Observed Meteorological Data”
- provided by University of Colorado Boulder
- contains both meteorological and hydrological data at the same temporal and geographical resolution

Changes

- we are now operating on a monthly data collection basis
- 1950-2013 only, so 2014-2024 was supplemented by a data projection

VIC:
The VIC (Variable Infiltration Capacity) model is a land surface and hydrology model that simulates land-atmosphere fluxes, and water and energy balances at the land surface. It uses meteorological data, soil parameters, and vegetation parameters as input.



THE PROJECTED DATA

01 LOCA VIC Runs

- provided by Scripps Institute of Oceanography
- provides the exact same complete variables as the historical VIC data, also on a monthly basis

Details

- 4 Predictive models with different assumptions
  - HadGEM2-ES: A “warmer/drier” simulation
  - CanESM2: An “average” simulation
  - CNRM-CM5: A “cooler/wetter” simulation
  - MIROC5: A “dissimilar” simulation that is most unlike the other three, to produce maximal coverage of possible future climate conditions
- RCP Emissions Trajectory
  - 4.5: Emission peak ~2040 and then decline
  - 8.5: Emission rise through 2050 and plateau around 2100

13

LOCA VIC = localized constructed analogs + variable infiltration campacity

LOCA:
The LOCA (Localized Constructed Analogs) technique is a statistical downscaling method that uses observed data to create high-resolution climate projections. It is used to correct biases in global climate model simulations and downscale the data to a finer spatial resolution.
VIC:
The VIC (Variable Infiltration Capacity) model is a land surface and hydrology model that simulates land-atmosphere fluxes, and water and energy balances at the land surface. It uses meteorological data, soil parameters, and vegetation parameters as input.

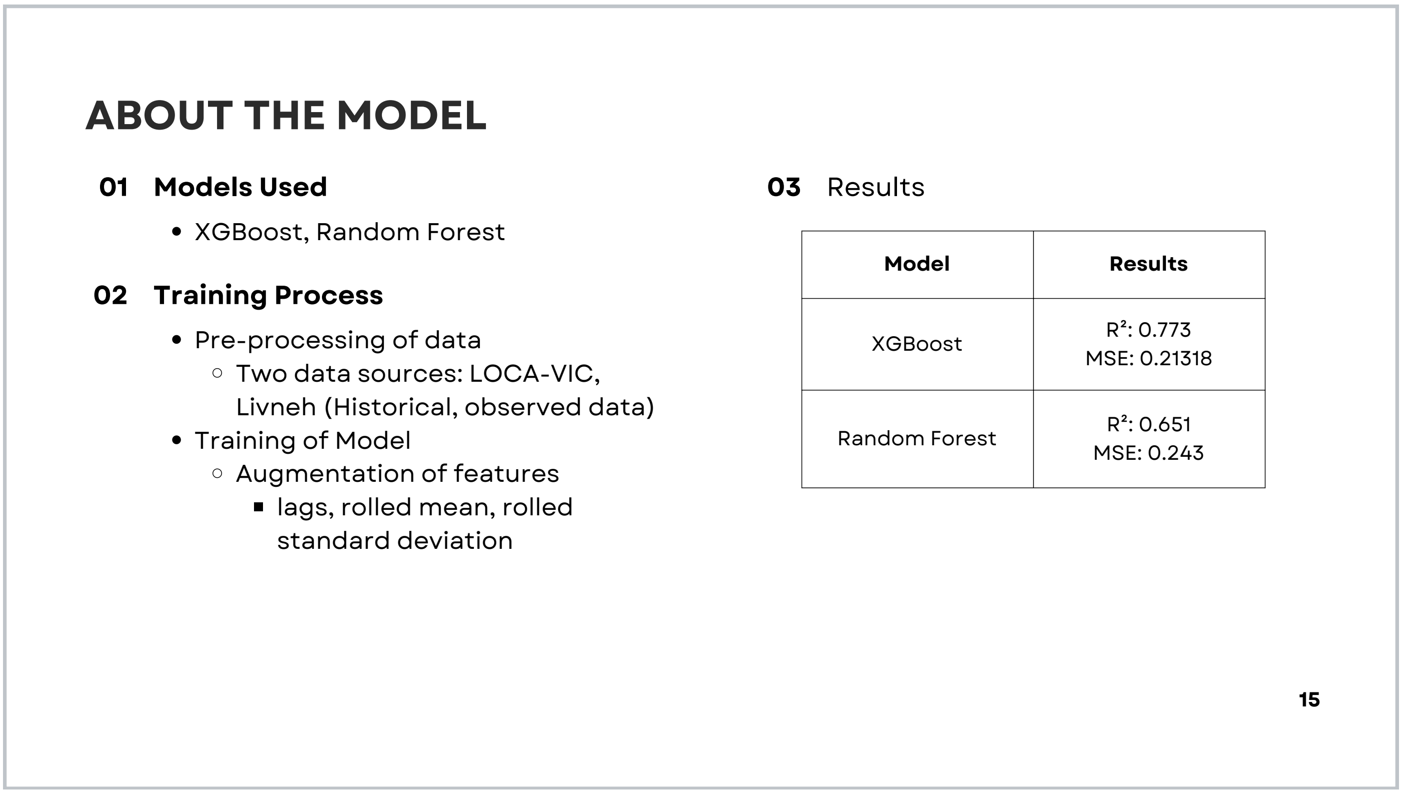
4.5 = 4.5 watts per meter squared of radiative forcing in the year 2100 = medium warning

8.5 = 8.5 watts per meter squared of radioactive forcing in the year 2100 = high warning

An emissions scenario is a representation of future greenhouse gas emissions and resulting atmospheric concentrations through time. An emissions scenario illustrates a plausible future so that climate projections for that emissions scenario can be generated, used to inform analysis and decision-making, and compared to other scenarios.

BUILDING THE PREDICTIVE MODEL

Actually, several.



ABOUT THE MODEL

01 Models Used

- XGBoost, Random Forest

02 Training Process

- Pre-processing of data
  - Two data sources: LOCA-VIC, Livneh (Historical, observed data)
- Training of Model
  - Augmentation of features
    - lags, rolled mean, rolled standard deviation

03 Results

Model	Results
XGBoost	R^2: 0.773 MSE: 0.21318
Random Forest	R^2: 0.651 MSE: 0.243

15

General model setup:
- D0-D4 categories = true labels, predict on the 8 variables

Initial process (different models, daily data, getting negative f1 and r^2 values)

Looking into "standardized" data
- less data conversion that loses precision over time
- monthly data -> account for lag AND OTHER VARS

2 models
- random forest
- xgboost

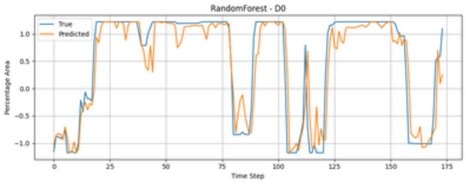
Livneh --> historical, observed data
Loca-vic --> predictive data using four different models. We're using HadGEM2-ES / warm-dry values/climate data predictive values to train for data from 2014-2024 and augmented

used xgboost and random forest to adapt to the non-linearity of climate data and to also predict multiple values, which is the D0-D4 values for each month given input features. Additionally, Random Forest and XGBoost can work with less data, and given we have limited data to use to train, we felt these models were appropriate.

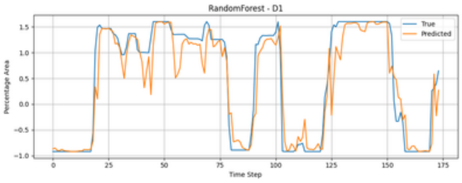
We pre-processed the data so that we had values from each month instead of daily, as the predicted values we used for 2014-2024 were monthly. We then decided to augment features by taking lag values and rolled mean and standard deviation values of d0-d4 to take into account the most recent values of d0-d4 as they do affect the current levels. For these values, we created lag values from 1 month ago, 2 months ago, and 3 months ago and rolled mean and standard deviation from the past seven months.

Our models performed quite well with average r^2 value of... and mse value for xgboost and average r^2 value of ... and mse value for random forest.

RF RESULTS

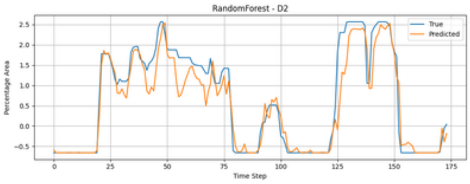


MSE: 0.2069,
R²: 0.8619

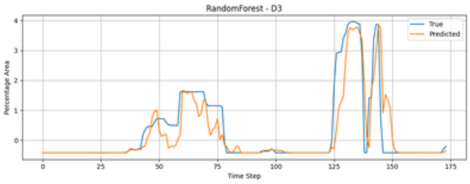


MSE: 0.2085,
R²: 0.8868

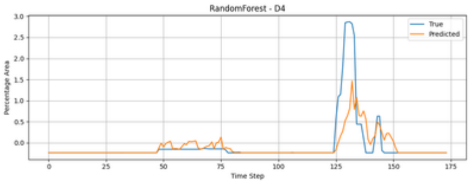
MSE: 0.2336,
R²: 0.8984



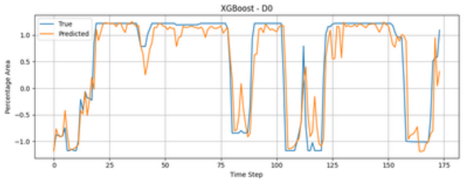
MSE: 0.2835,
R²: 0.7385



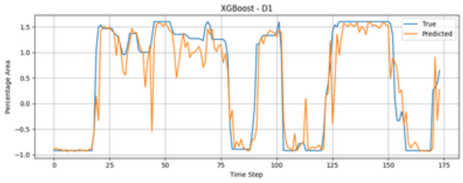
MSE: 0.1334,
R²: 0.4794



XGBOOST RESULTS

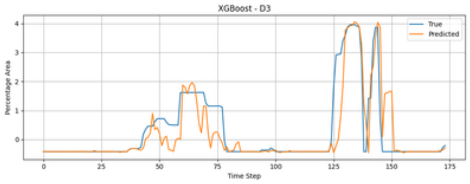


MSE: 0.2126,
R^2: 0.8543

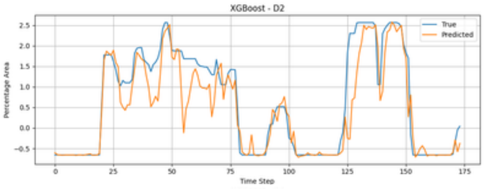


MSE: 0.2241,
R^2: 0.8560

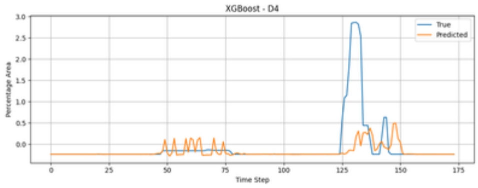
MSE: 0.2884,
R^2: 0.8278



MSE: 0.3245,
R^2: 0.6535

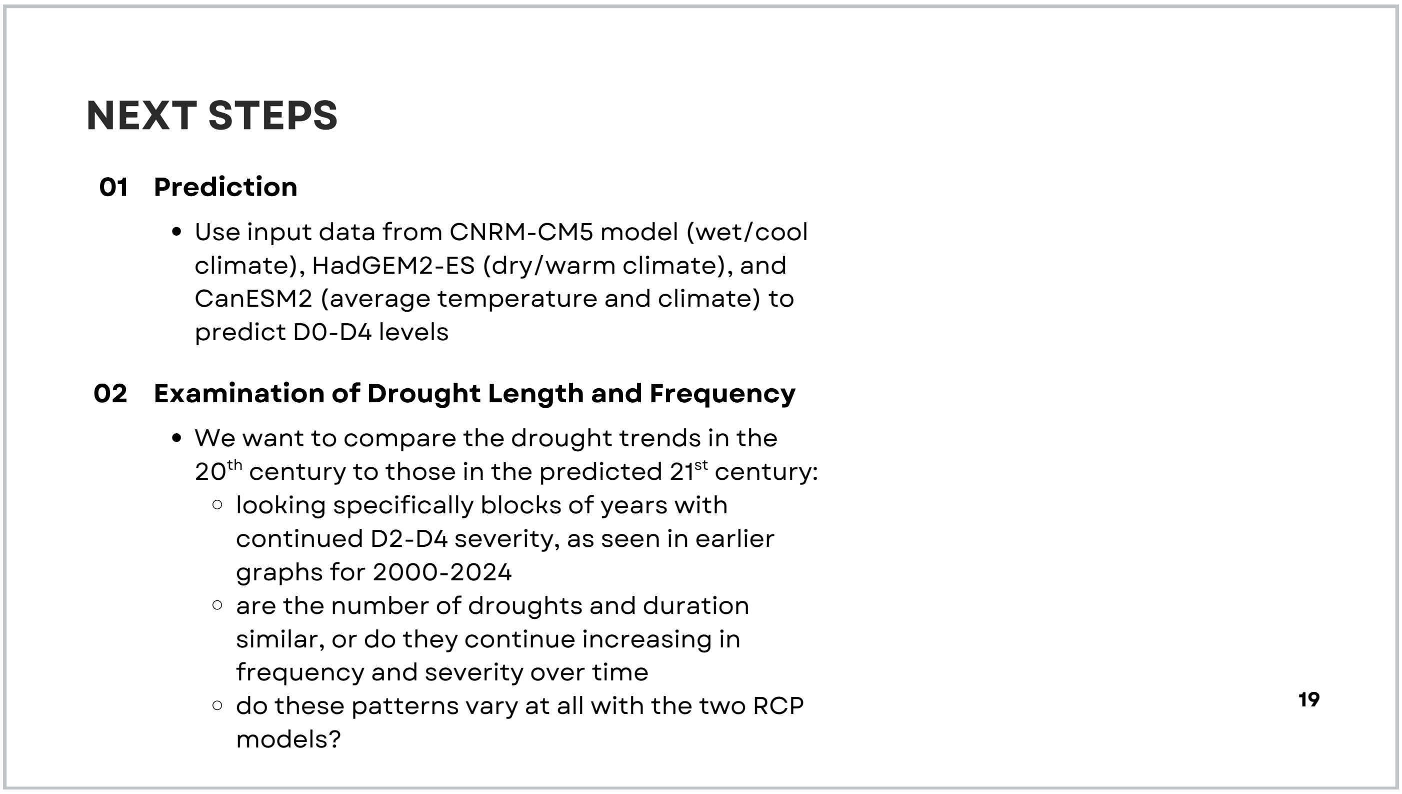


MSE: 0.1676,
R^2: 0.0631



MODELING THE FUTURE

For several scenarios.



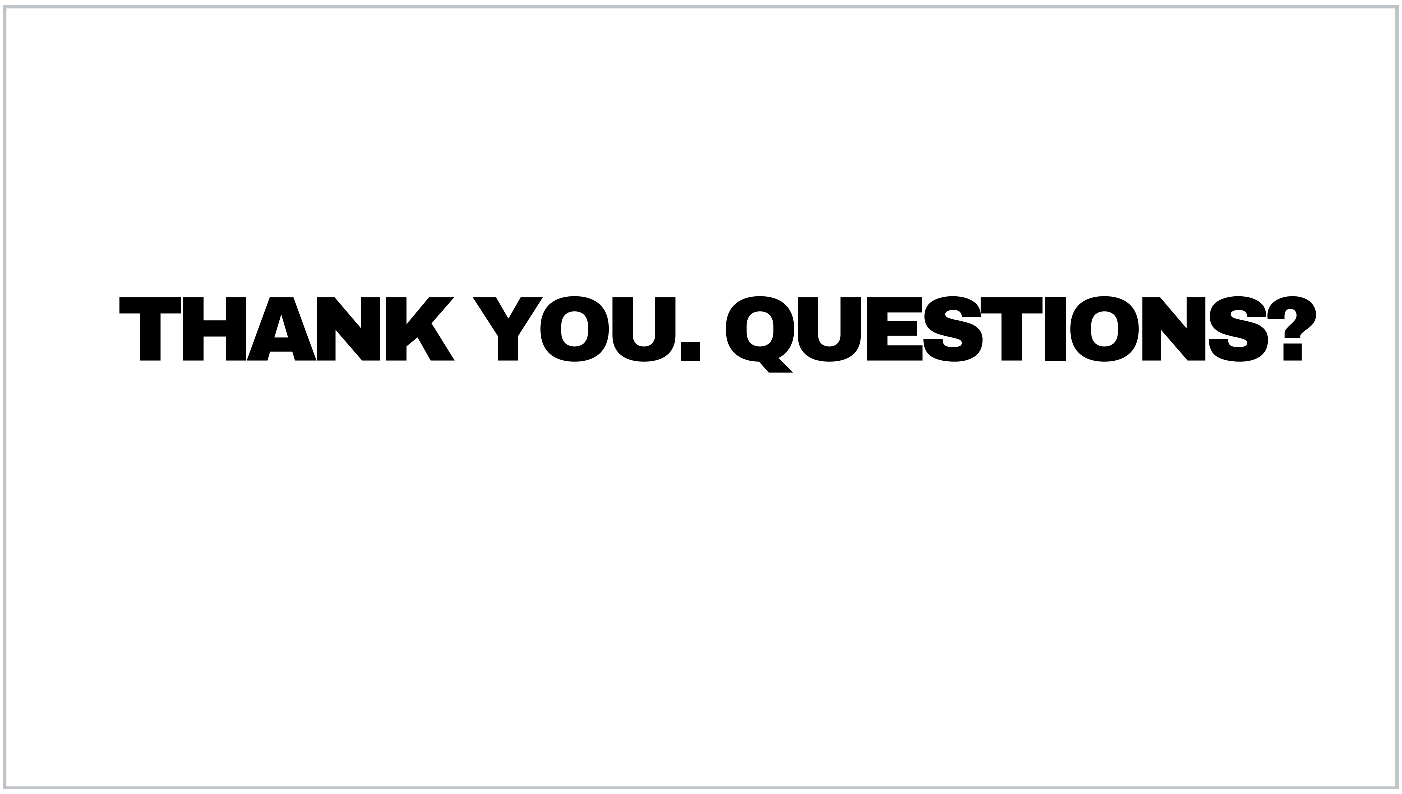
NEXT STEPS

01 Prediction

- Use input data from CNRM-CM5 model (wet/cool climate), HadGEM2-ES (dry/warm climate), and CanESM2 (average temperature and climate) to predict D0-D4 levels

02 Examination of Drought Length and Frequency

- We want to compare the drought trends in the 20th century to those in the predicted 21st century:
  - looking specifically blocks of years with continued D2-D4 severity, as seen in earlier graphs for 2000-2024
  - are the number of droughts and duration similar, or do they continue increasing in frequency and severity over time
  - do these patterns vary at all with the two RCP models?



THANK YOU. QUESTIONS?