Ladies
Melbourne

Random Forests, Climate Extreme Events and Food Production

Elisabeth Vogel, PhD student
Australian German Climate & Energy College
ARC Centre of Excellence on Climate System Science

25/10/2017

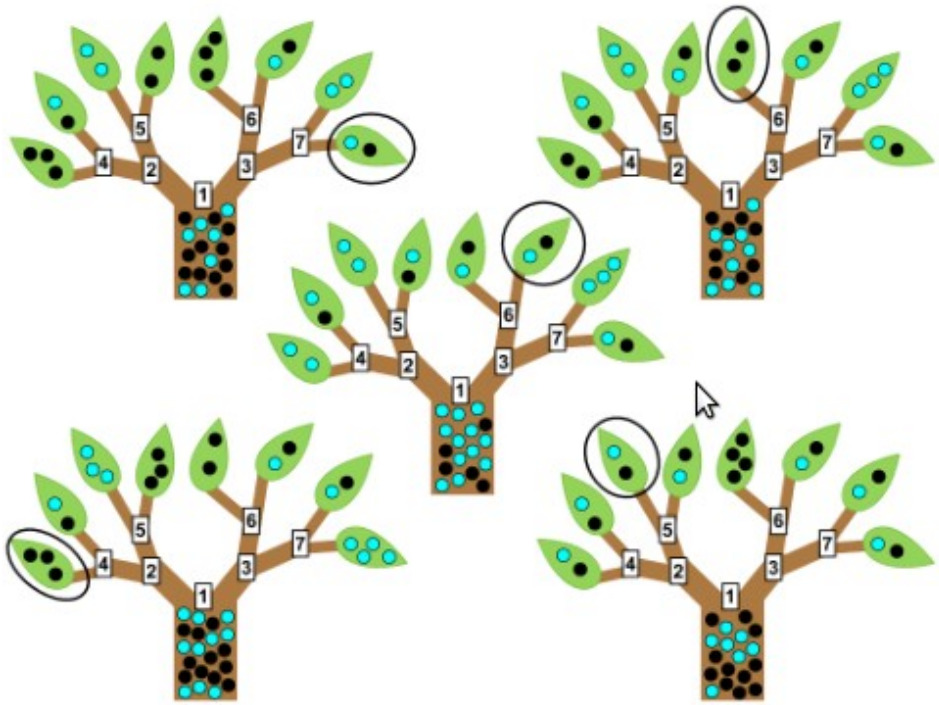


Image source: mechanicalforex.com/wp-content/uploads/2016/08/Selection_999256.jpg

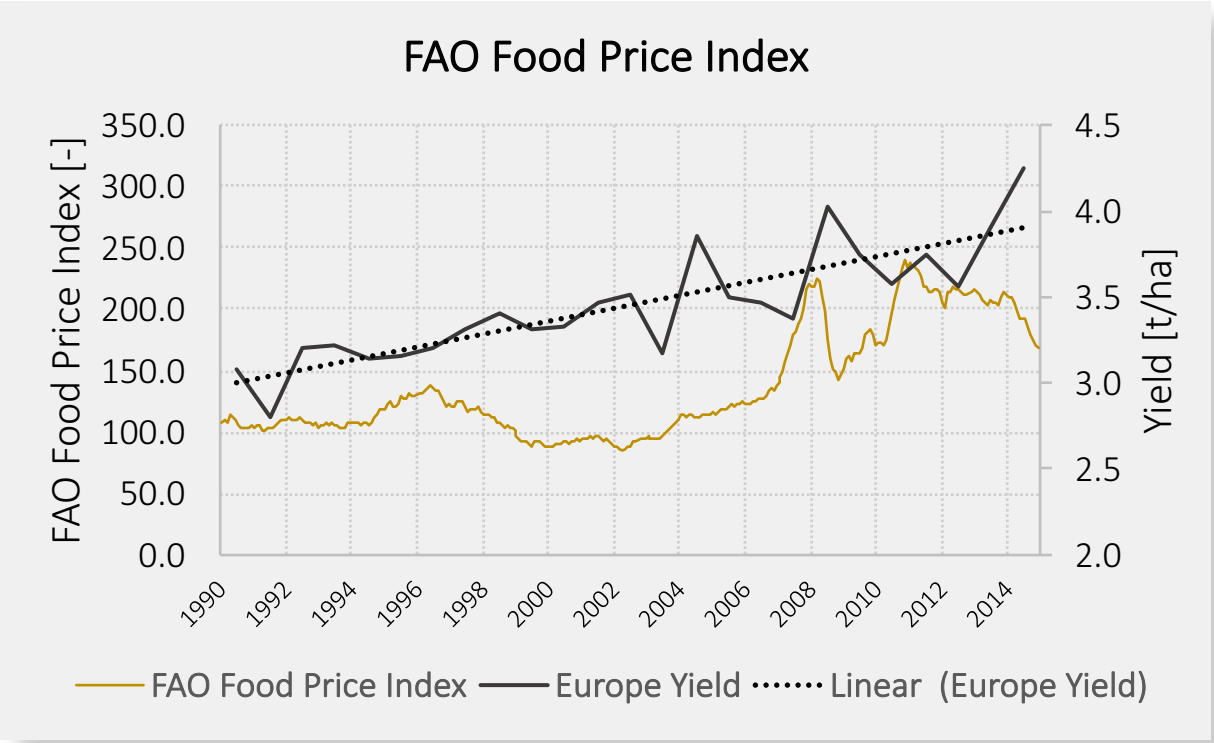
About me

- MSc in Environmental Engineering, Technical University Berlin
    - Master thesis: The temporal and spatial variability of soil respiration in boreal forests - A case study of Norunda forest, Central Sweden
- since 2014: PhD student at the Australian-German Climate & Energy College, The University of Melbourne
    - **Research topic:** The impact of climate extreme events on global agricultural yields
    - **Research interests:** Climate change impacts, extreme events, agriculture, ecosystems; Statistical and machine learning tools for data analysis and predictions

In a random forest somewhere in Sweden

Why do we care about climate extremes and agriculture?

- Climate change increases the risks of certain types of extreme events, such as droughts and heatwaves
- Agricultural sector particularly vulnerable to extreme weather events as it depends on weather conditions during the growing season
- Production shocks → Price spikes that have impacts not only locally, but worldwide
  Example: 2007 European heatwave; 2010 Russian heatwave
- Important to get a better understanding of the impact of extreme events on crop yields to adapt the global food system



Source: FAOSTAT

Overview

- Part 1: Random Forests
  - What are Random Forests?
  - Advantages / Disadvantages
  - How to set up Random Forests in R

- Part 2: Can we use Random Forests to analyse the impacts of climate extreme events on global agricultural yields?


Part 1: Random Forests

What is a “Random Forest”?

- A machine learning algorithm based on an ensemble of decision trees → used for predictive modelling
- For classification and regression problems
  - **classification:** predicting to which category a new observation belongs, e.g. medicine: high or low risk of diabetes
  - **regression:** predicting continuous values
- **Supervised learning method:**
  - Takes a training dataset of predictor and predicted variables and identifies relationships in the data → these relationships are then used to predict the outcome for new observations or samples
- Described in: Breiman (2001)

A single decision tree

Netflix suggesting movies:
Will I like this movie?

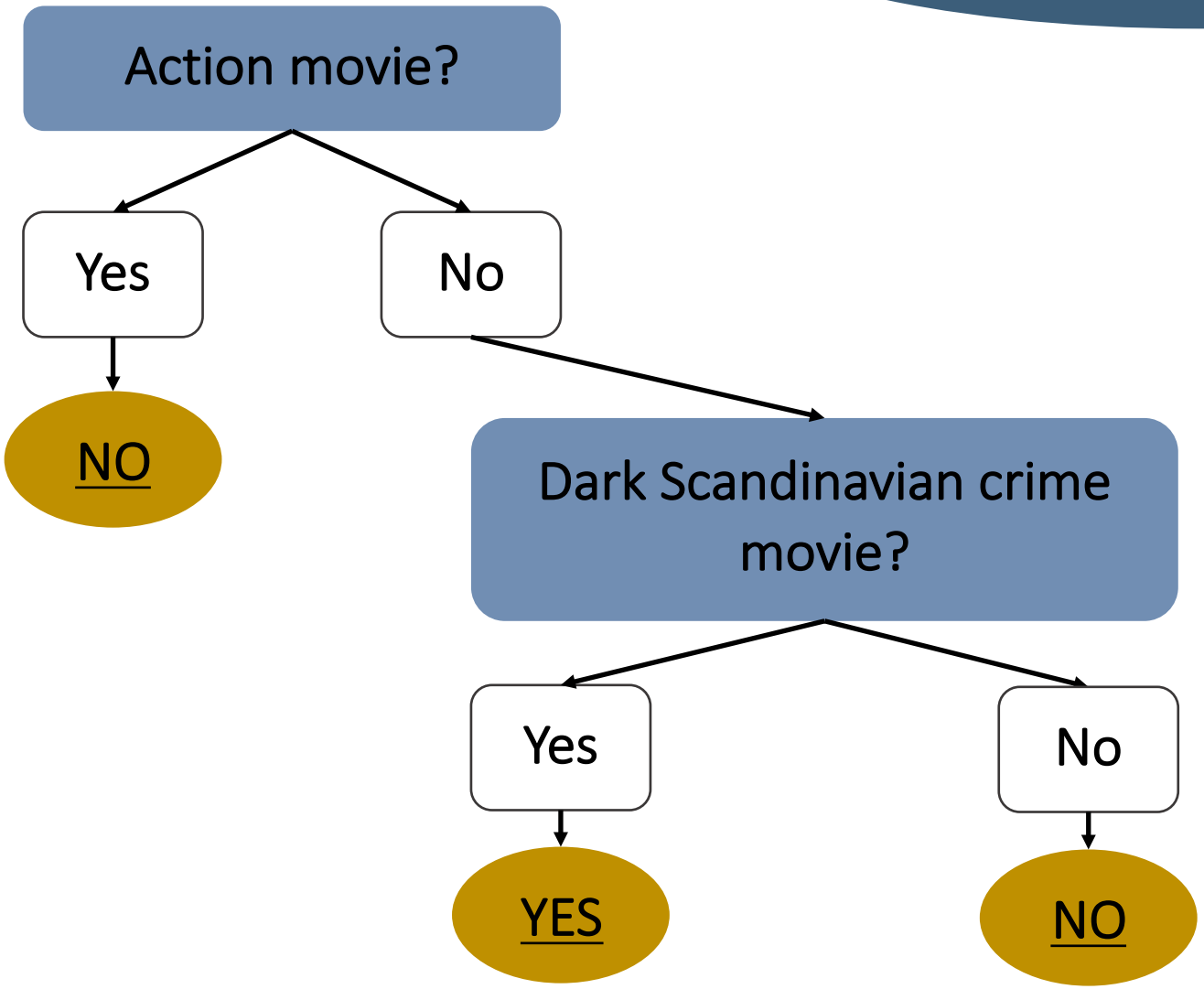
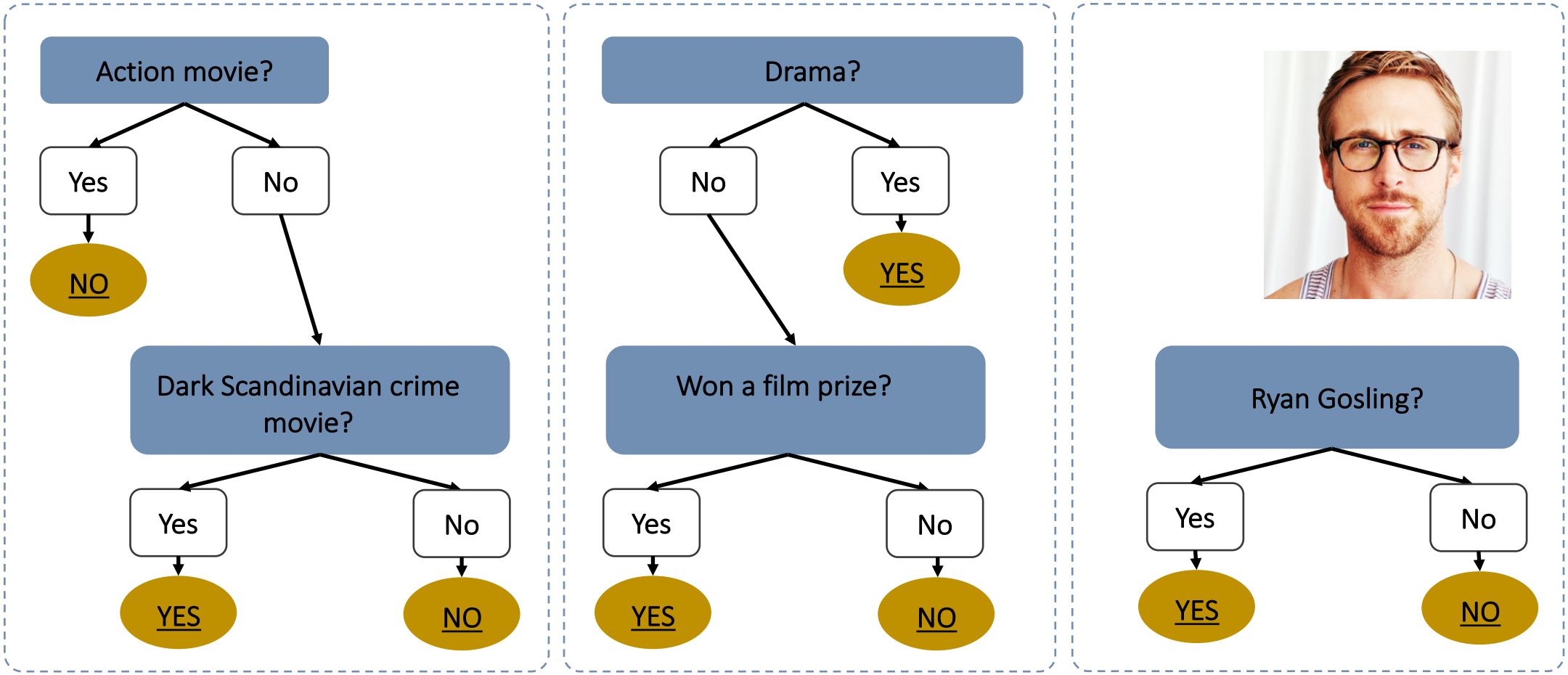


Image source: bonocle.files.wordpress.com/2012/01/tumblr_ls3tb9dqfp1qf8r8uo1_500.png

Combination of decision trees

Netflix suggesting movies:
Will I like this movie?



Decision trees

- Biggest disadvantage is that they tend to overfit the data, especially when they are very deep
- Trade-off between bias and variance:
  - Small trees: high bias, low variance
  - More complex trees: low bias, but high variance
- May therefore have poor performance on new observations
- Can be overcome by growing a large number of independent decision trees based on random sub-samples of the dataset → Random forest

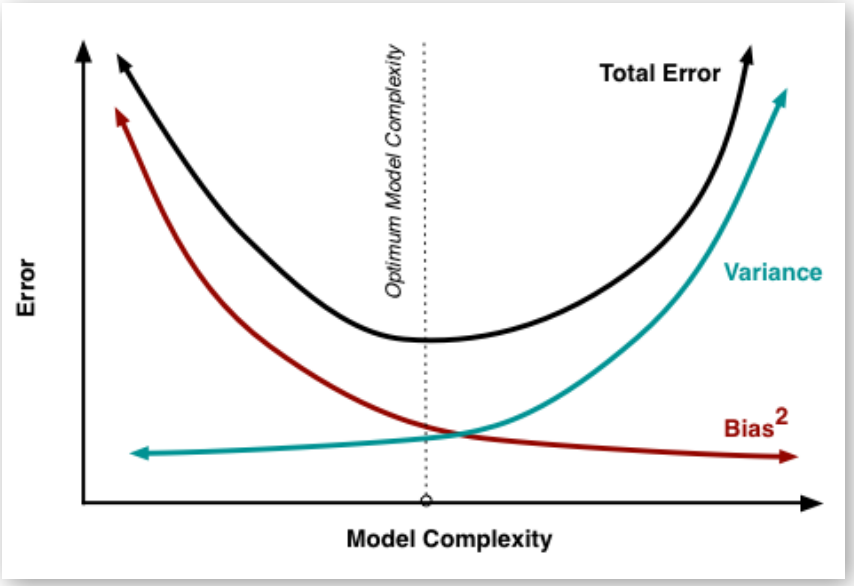


Image source: scott.fortmann-roe.com/docs/docs/BiasVariance/biasvariance.png

Random forests

Source: https://www.kaggle.com/c/titanic/data

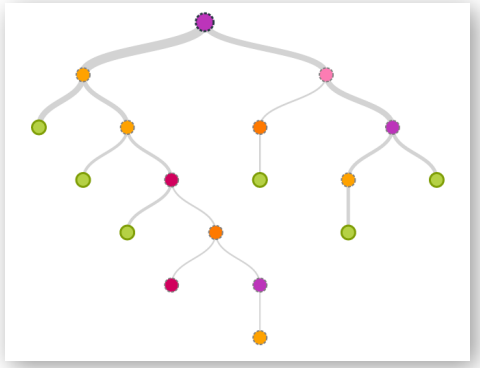
	PassengerId	Survived	Pclass	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
1	1	0	3	male	22.00	1	0	A/5 21171	7.2500		S
2	2	1	1	female	38.00	1	0	PC 17599	71.2833	C85	C
3	3	1	3	female	26.00	0	0	STON/O2. 3101282	7.9250		S
4	4	1	1	female	35.00	1	0	113803	53.1000	C123	S
5	5	0	3	male	35.00	0	0	373450	8.0500		S
6	6	0	3	male	NA	0	0	330877	8.4583		Q
7	7	0	1	male	54.00	0	0	17463	51.8625	E46	S
8	8	0	3	male	2.00	3	1	349909	21.0750		S
9	9	1	3	female	27.00	0	2	347742	11.1333		S
10	10	1	2	female	14.00	1	0	237736	30.0708		C
11	11	1	3	female	4.00	1	1	PP 9549	16.7000	G6	S
12	12	1	1	female	58.00	0	0	113783	26.5500	C103	S
13	13	0	3	male	20.00	0	0	A/5. 2151	8.0500		S
14	14	0	3	male	39.00	1	5	347082	31.2750		S
15	15	0	3	female	14.00	0	0	350406	7.8542		S
16	16	1	2	female	55.00	0	0	248706	16.0000		S

Data set:

- *K* predictor variables (independent variables, features)
- 1 predicted variable (dependent variable, label)
- *N* observations (samples)

For every decision tree d:

i. randomly select a subset of *n* observations (with replacement)
ii. create an optimal decision rules to sub-divide the data
iii. at every decision split: only use a randomly selected subset of *k* predictors to choose from for decision rule
iv. repeat iii) until specified tree depth is reached

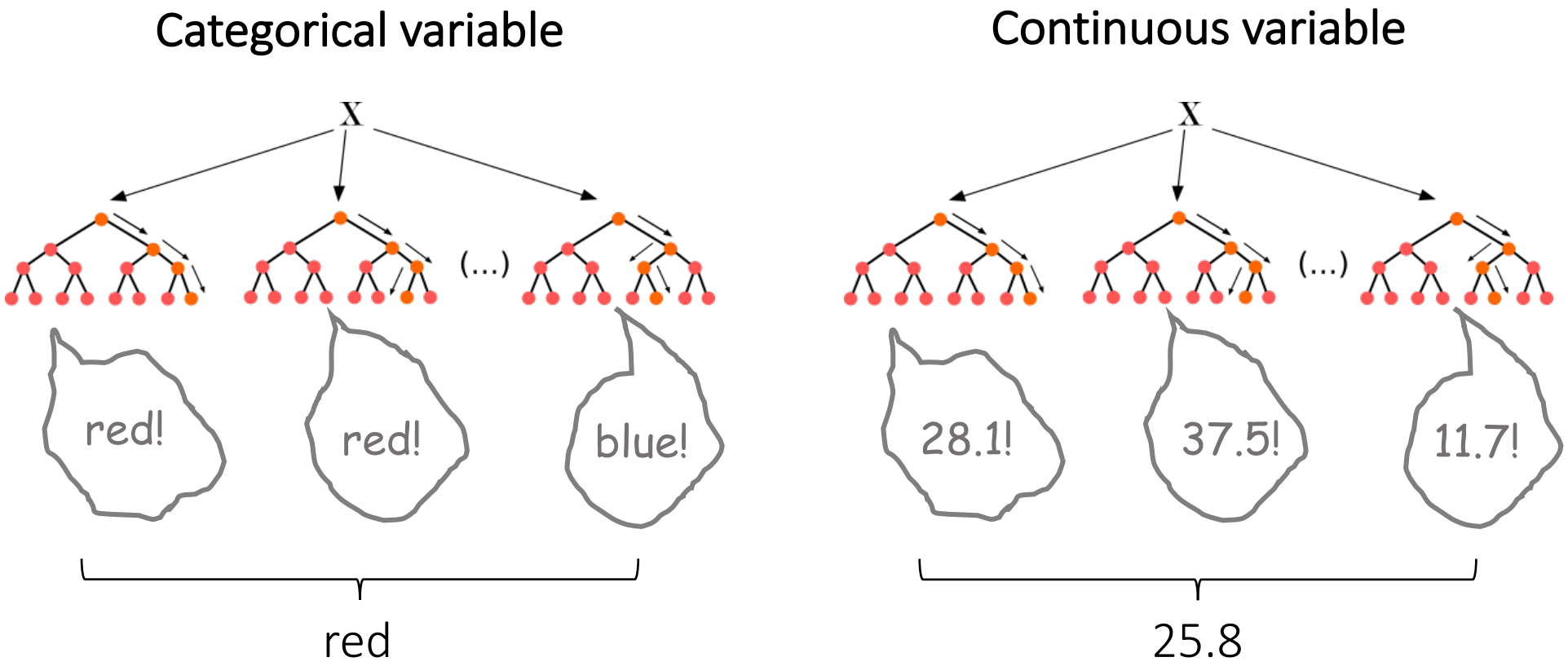


Random forest

- grow *D* decision trees

Random forests

Predictions: using majority vote (categorical variables) or mean (continuous variables)



Advantages, drawbacks and applications

Advantages:

- Good predictive skill
- Can be used for predictions of continuous, binary and categorical variables
- Almost no predictor preparation:
  - can use continuous or categorical variables
  - no linearity or normality requirements
  - can be used to impute missing values
- Robust to collinearities in the data
- Selects most important features and ignores other features

Drawbacks:

- “black box”, hard to interpret (diagnostic plots can help visualise some of the patterns)
- Can become large in size and slow to train compared to linear models

Examples of applications:

- Land use classification from satellite images
- Prediction of biological properties of molecules
- Assessment of credit risks

Creating Random Forests in R

```r
# Install and load randomForest package install.packages('randomForest')
library(randomForest)

# Install and load MASS package for sample datasets install.packages('MASS')
library(MASS)
```

```r
# Boston dataset: Housing Values in Suburbs of Boston
str(Boston)
```

```
FALSE 'data.frame': 506 obs. of  14 variables:
FALSE  $ crim    : num  0.00632 0.02731 0.02729 0.03237 0.06905 ...
FALSE  $ zn      : num  18 0 0 0 0 0 12.5 12.5 12.5 12.5 ...
FALSE  $ indus   : num  2.31 7.07 7.07 2.18 2.18 2.18 7.87 7.87 7.87 7.87 ...
FALSE  $ chas    : int  0 0 0 0 0 0 0 0 0 0 ...
FALSE  $ nox     : num  0.538 0.469 0.469 0.458 0.458 0.458 0.524 0.524 0.524 0.524 ...
FALSE  $ rm      : num  6.58 6.42 7.18 7 7.15 ...
FALSE  $ age     : num  65.2 78.9 61.1 45.8 54.2 58.7 66.6 96.1 100 85.9 ...
FALSE  $ dis     : num  4.09 4.97 4.97 6.06 6.06 ...
FALSE  $ rad     : int  1 2 2 3 3 3 5 5 5 5 ...
FALSE  $ tax     : num  296 242 242 222 222 222 311 311 311 311 ...
FALSE  $ ptratio: num  15.3 17.8 17.8 18.7 18.7 18.7 15.2 15.2 15.2 15.2 ...
FALSE  $ black   : num  397 397 393 395 397 ...
FALSE  $ lstat   : num  4.98 9.14 4.03 2.94 5.33 ...
FALSE  $ medv    : num  24 21.6 34.7 33.4 36.2 28.7 22.9 27.1 16.5 18.9 ...
```

based on: T. Hastie and R. Tibshirani Online Class on Statistical Learning

Creating Random Forests in R and skill assessment

```
# Define a training subset (for cross-validation)
set.seed(101)
train = sample(1:nrow(Boston), 300)
```

```
# Create random forest for predicting median house values
rf.boston = randomForest(medv ~ ., data = Boston, subset = train)
rf.boston
```

```
##
## Call:
##  randomForest(formula = medv ~ ., data = Boston, subset = train)
##                Type of random forest: regression
##                      Number of trees: 500
## No. of variables tried at each split: 4
##
##           Mean of squared residuals: 12.34243
##                     % Var explained: 85.09
```

```
# Assessing the skill Out-of-bag mean-squared error (MSE)
rf.boston$mse[length(rf.boston$mse)]
```

```
## [1] 12.34243
```

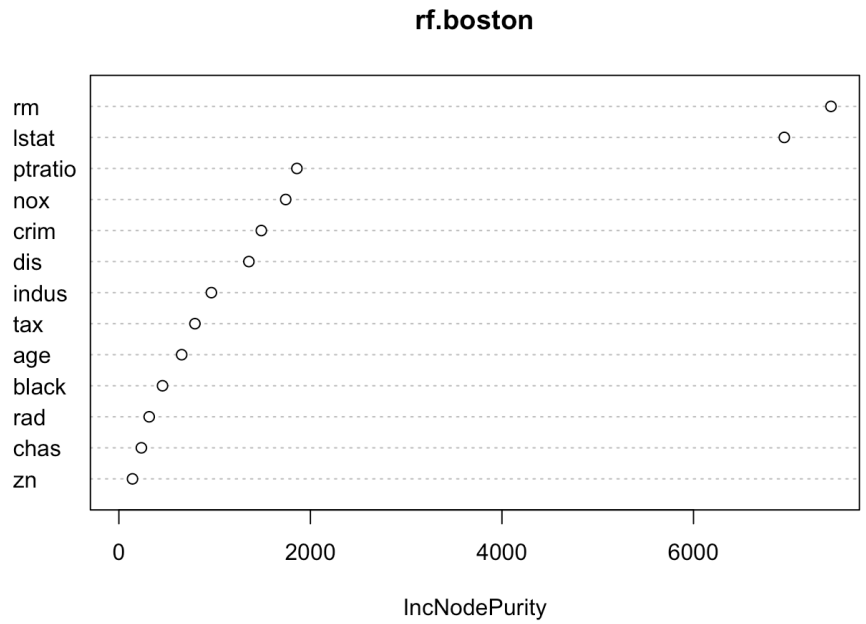
```
# Out-of-bag R-squared (Rsq)
rf.boston$rsq[length(rf.boston$rsq)]
```

```
## [1] 0.8508772
```

based

Visualisation of the model

```
# Variable importance plots
varImpPlot(rf.boston)
```

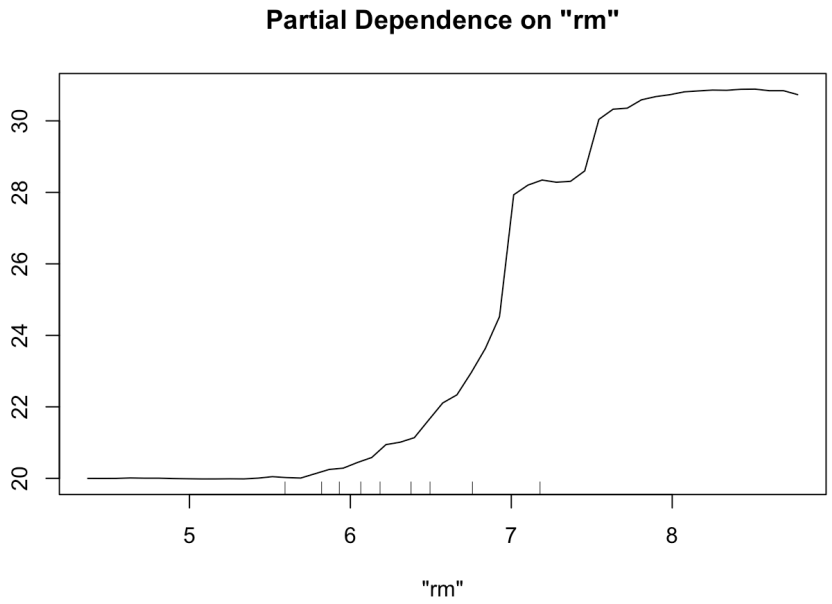


- **Variable importance plots** show the average reduction in performance after randomly perturbing each predictor at every decision split
- Two possibilities:
  - increase in mean squared error (%IncMSE)
  - mean decrease in Gini coefficient (%IncNodePurity)
- The greater the variable importance value, the more important the variable with respect to the predictive capacity of the model
- Variables ranked by importance: highest to lowest

based on: T. Hastie and R. Tibshirani Online Class on Statistical Learning

Visualisation of the model

```
# Partial dependence plots
partialPlot(rf.boston, x.var = "rm", pred.data = Boston[train, ])
```



Partial dependence plots:

- Show the marginal effect (positive or negative) of a predictor variable
- How does the predictor influence the outcome, if all other variables are accounted for?

based on: T. Hastie and R. Tibshirani Online Class on Statistical Learning

Part 2: Impacts of Climate Extreme Events on Global Agricultural Yields

Research approach

Main challenges:

- Data availability at appropriate spatial and temporal scale
- Length of available time series

- Collinearity between predictor variables
- Complex interactions between climate variables and yield variable
- Non-linear effects

Research approach:

- Use a global gridded crop yield data base 47 years of yield data for wheat, maize, rice
- combined with global gridded dataset on extreme event indicators (HADEX2) and global, gridded historical climate observation data (CRU TS 3.23)

- Application of **random forest algorithm** that is robust to collinearities between predictor variables and that can capture complex interactions between climate variables and agricultural productivity

Research approach

Main challenges:

- Data availability at appropriate spatial and temporal scale
- Length of available time series

- Collinearity between predictor variables
- Complex interactions between climate variables and yield variable
- Non-linear effects

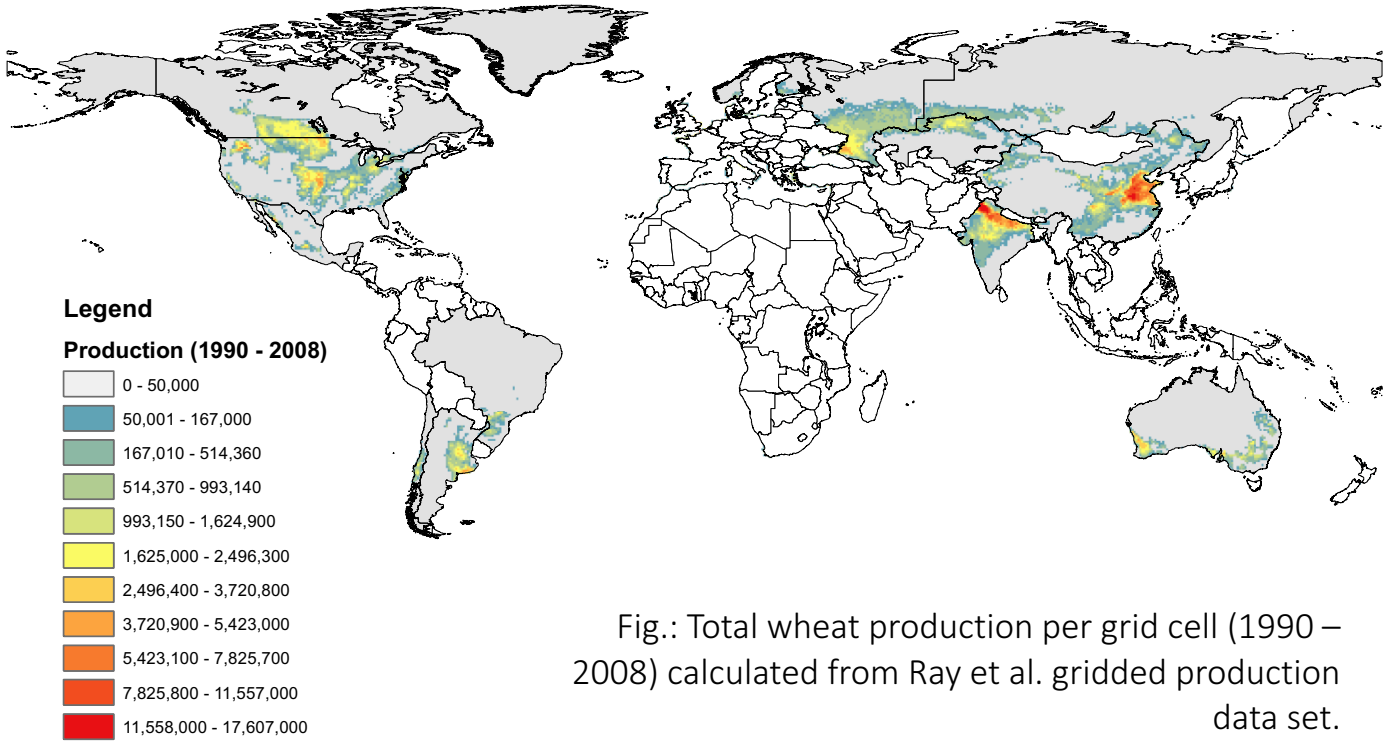
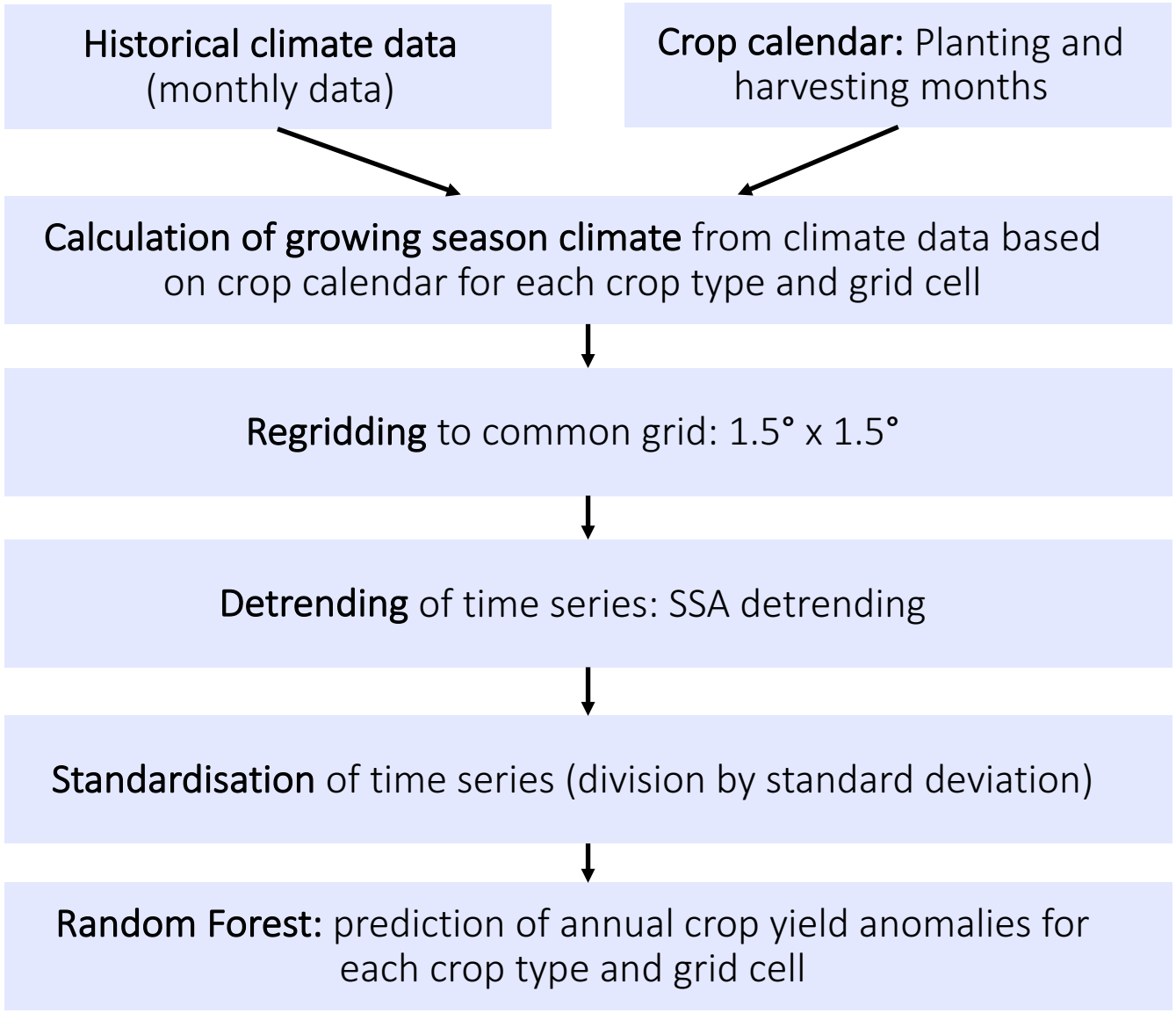


Fig.: Total wheat production per grid cell (1990 – 2008) calculated from Ray et al. gridded production data set.

Methodology – Data preparation



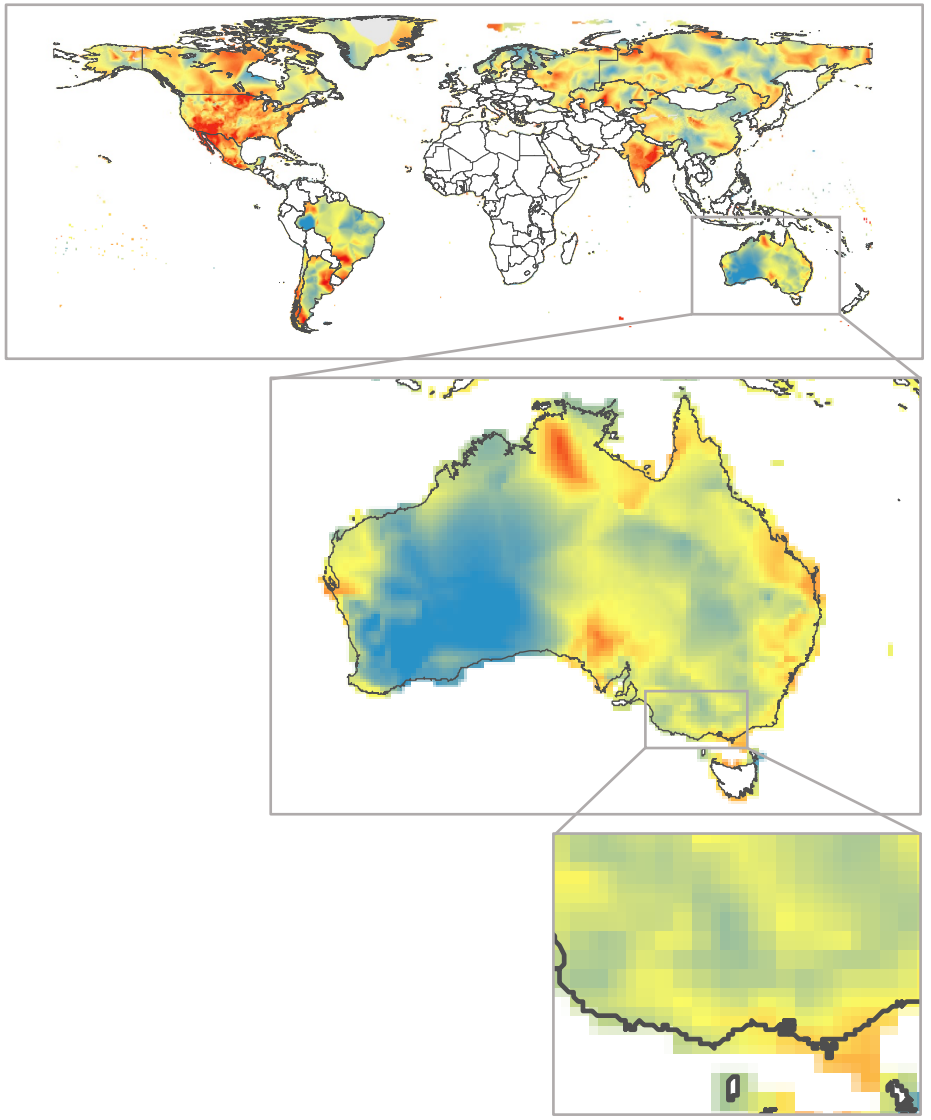
Cross-validation:

- 5 iterations: 80% training years, 20% test years
- Prediction of yield anomalies for test years → combined after 5 iterations to one time series (out-of-sample predictions)
- Calculation of skill metrics for OOS-predictions (R2, RMSE)

Methodology – Predictors

- Mean climate during the growing season:
  - mean temperature
  - mean precipitation
  - mean daily temperature range

- Extreme event indicators
  - maximum temperature
  - minimum temperature
  - frequency of unusually warm days (> 90th percentile)
  - frequency of unusually cold nights (< 10th percentile)
  - frost day frequency
  - maximum 5-day precipitation intensity
  - drought indicator: Standardised Precipitation Index (SPI-6)



NetCDF file format

- “Network Common Data Form”
- Standard file format in the atmospheric, ocean and climate sciences for storing gridded data, e.g.
  - climate model output
  - meteorological observations
- set of software libraries that can create, read and write netCDF files
- stores metadata inside the file (e.g. variable names, units)
- portable file format
- can be a steep learning curve and sometimes errors related to different netcdf versions / naming conventions etc.

- R libraries to work with netcdf files: RNetCDF, ncdf4 and others

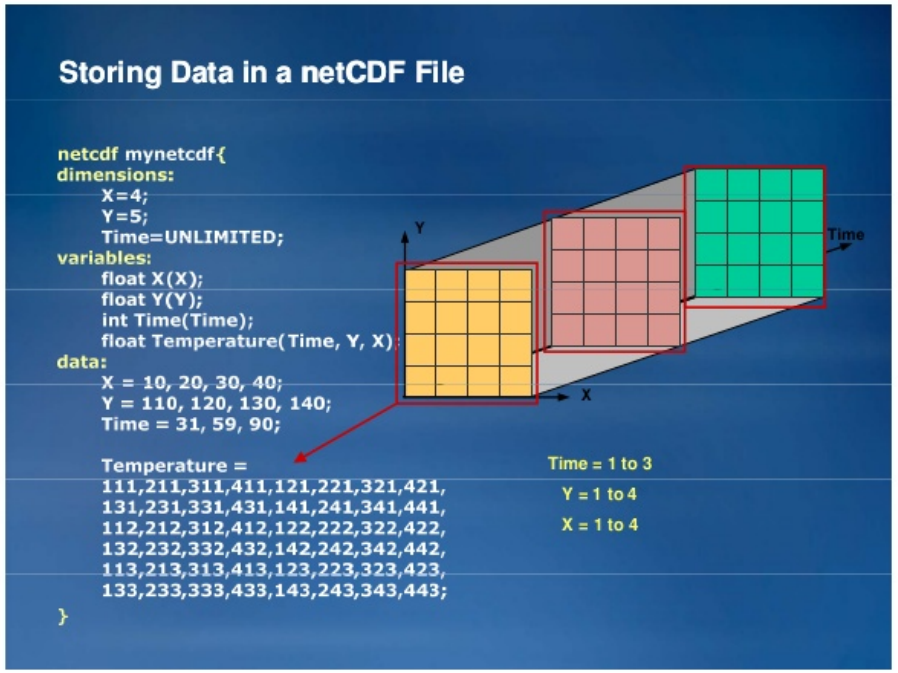
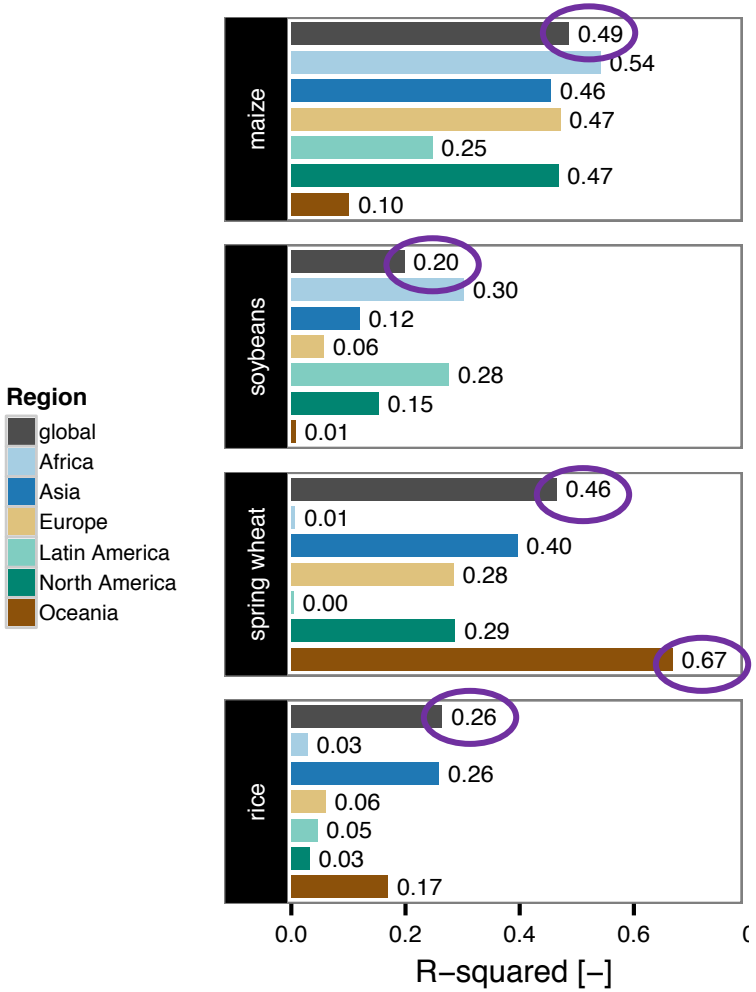


Image source: www.slideshare.net/HDFEOS/working-with-hdf-and-netcdf-data-in-arcgis-tools-and-case-studies

How much of the global and continental yield variability is explained by the random forest model?

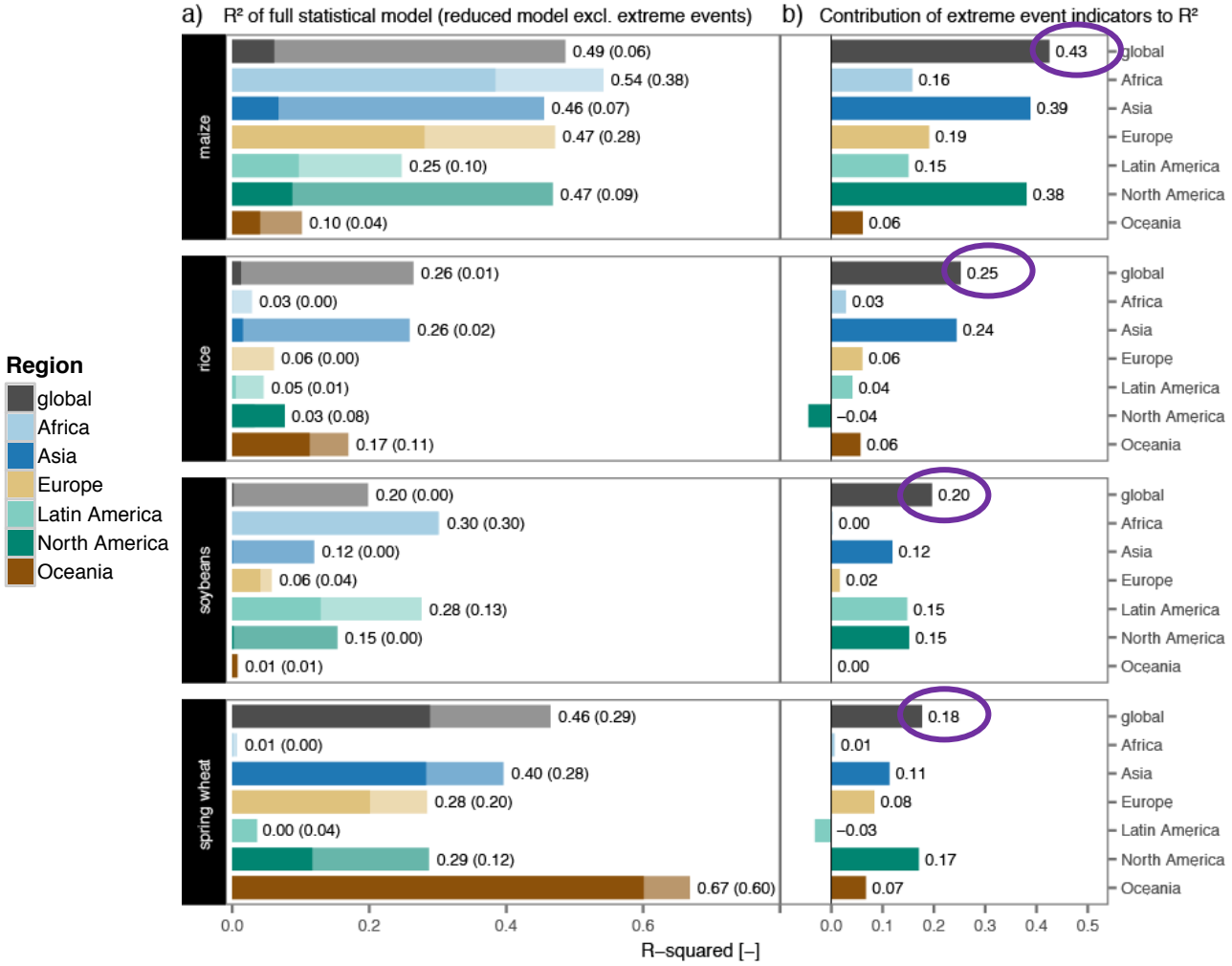
R2 value of correlation between actual and predicted yield anomalies, aggregated over continents and globally



- Our statistical model is able to explain about 50% of variability of detrended global maize and spring wheat yields
- Soybeans and rice: 20-25% of variance of anomalies explained
- Wheat fluctuations captured best in Australia: approx. 70%

What is the fraction of variability explained by extreme indicators?

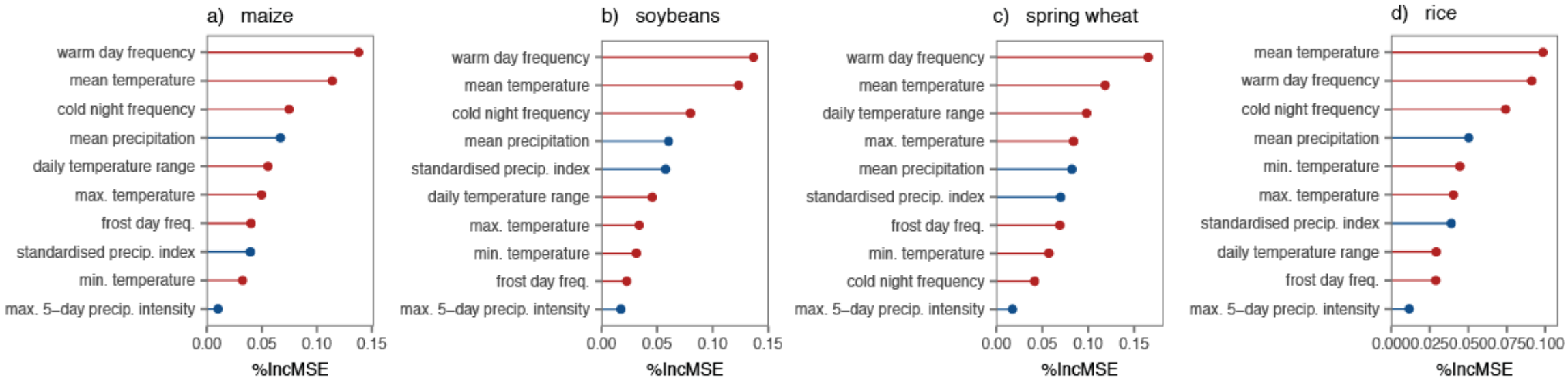
R2 value of correlation between actual and predicted yield anomalies, aggregated over continents and globally



- Two predictor sets:
  - **full model:** using all predictor variables, including mean climate and extreme indicators
  - **reduced model:** only precipitation and temperature
- Extreme event indicators increase explained variability considerably at the global scale as well as for certain regions
- Globally, increases of R^2:
  - Maize: 43%
  - Rice: 25%
  - Soybeans: 20%
  - Spring wheat: 18%

Which predictors have the greatest influence on crop yields? - Variable importance ranking

Variable importance plots for all area harvested, globally

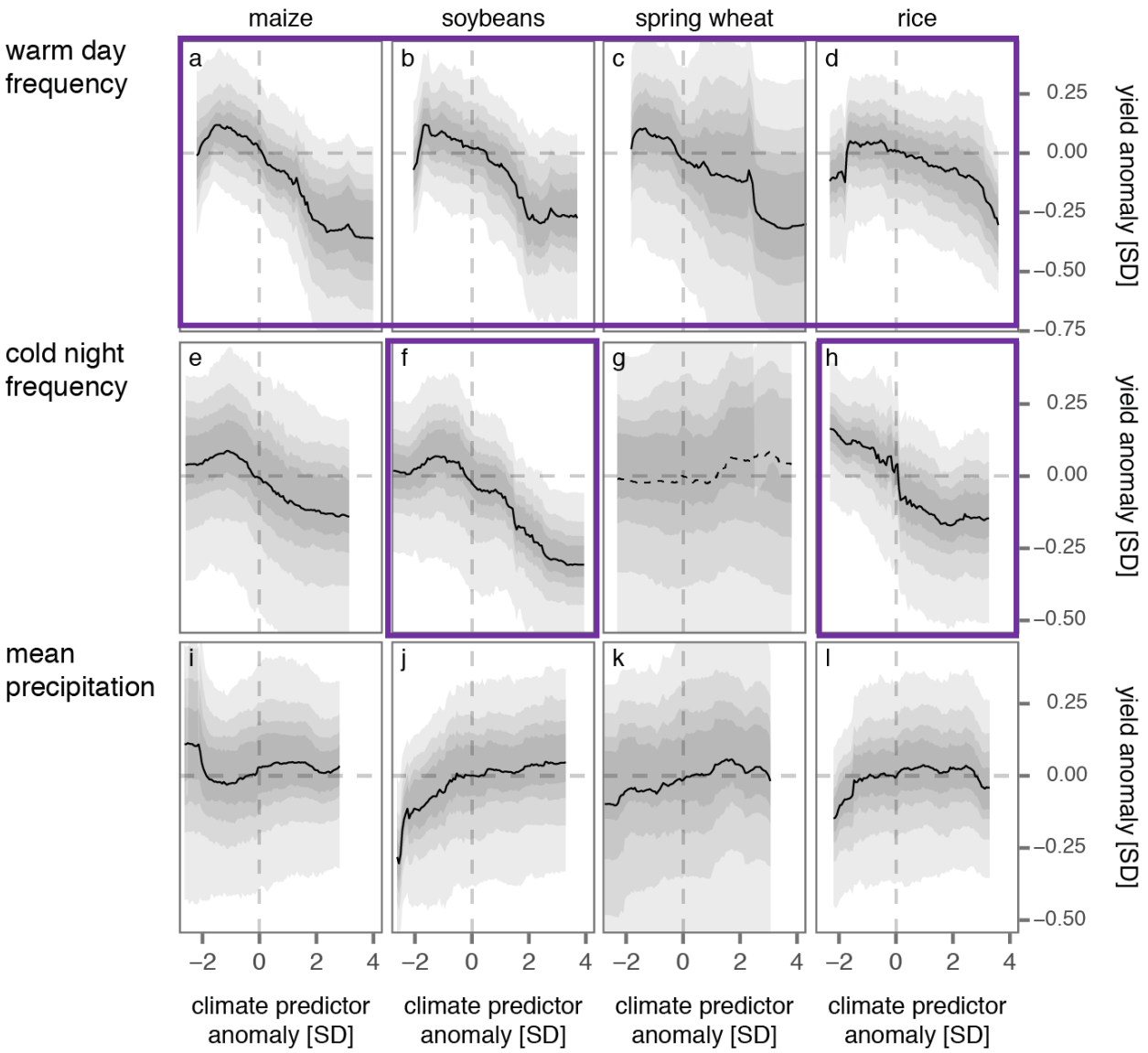


Blue: Rainfall-related predictors

Red: Temperature-related predictors

- Temperature-related indices are most important predictors for all crops
- Precipitation has comparatively low importance metrics
- Lowest: 5-day rainfall intensity (heavy precipitation index)

The marginal effect of predictors: partial dependence plots



- Clearly negative effect of increases in extreme warm days
- Positive effect of decreases in cold nights
- Association with precipitation less clear

Summary - some lessons I learned

- **Be aware of auto-correlation / data leakage:**
  for example, if datasets have different underlying spatial resolutions, the random forest will use the coarser dataset to predict neighboring grid cells due to spatial auto-correlation / data leakage

- **Skill metrics:**
  default skill metrics of random forest package: out-of-bag (OOB) error and R2 → usually very high values, better to subset the dataset manually and calculate own metrics

- **Partial dependence plots:**
  include uncertainty bands to avoid over-interpretation of the functional relationships

Summary – Random forests

- Random forests are a machine learning technique used for classification and prediction
- High accuracy and efficiency on large datasets
- Only little data preparation needed, can be used on different data types and to fill missing values
- “Black-box”, but visualisation tools allow to better understand the relative influence of each predictor

Thank you!

And Happy Birthday, R-Ladies! :)

Contact: elisabeth.vogel@climate-energy-college.org

References and additional resources

- Tutorial on Statistical Learning (including Random Forests): https://www.r-bloggers.com/in-depth-introduction-to-machine-learning-in-15-hours-of-expert-videos/
- Introduction to Random Forests by Breiman and Cutler: https://www.stat.berkeley.edu/~breiman/RandomForests/cc_home.htm
- Breiman (2001): http://oz.berkeley.edu/~breiman/randomforest2001.pdf
- Kaggle – "Learning Machine Learning from Disaster": https://www.kaggle.com/c/titanic
- NetCDF: http://geog.uoregon.edu/bartlein/courses/geog490/week04-netCDF.html