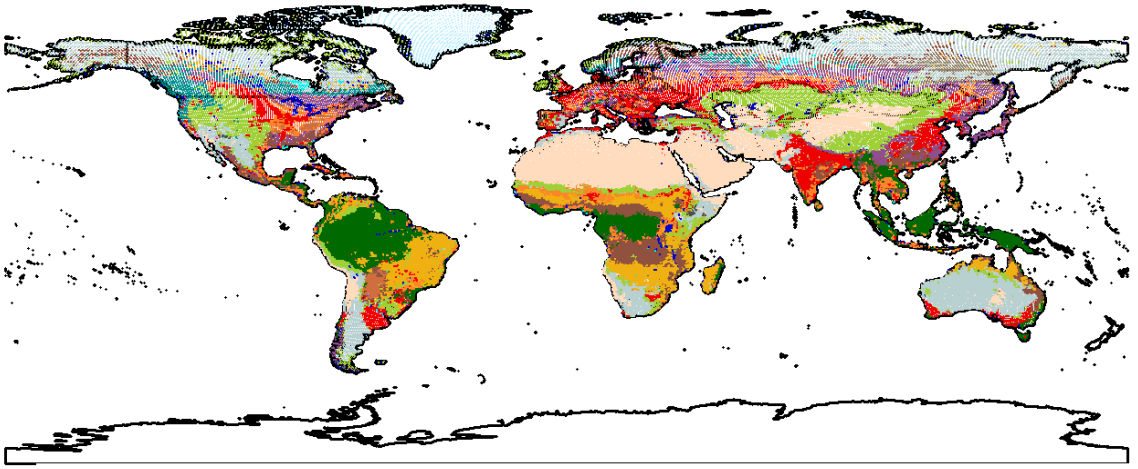
Multi-output prediction of global vegetation distribution with incomplete data

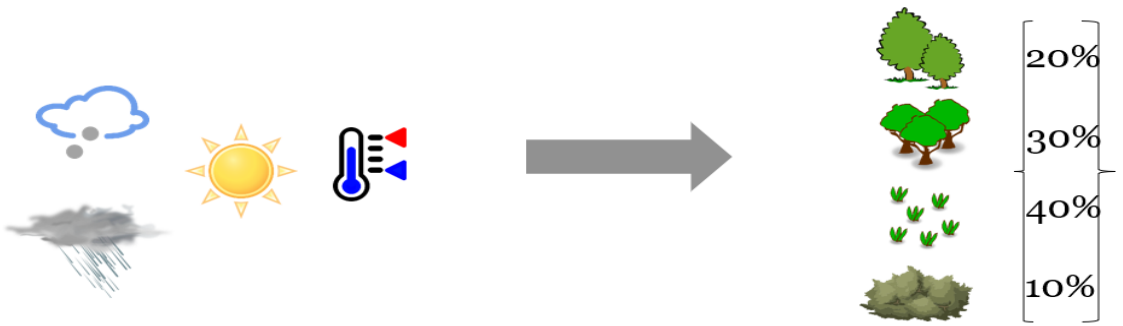
Rita Beigaite, University of Helsinki
Jesse Read, LIX, École Polytechnique
Indre Zliobaite, University of Helsinki

ARTEMISS workshop, July 17th, 2020

Introduction and Aim

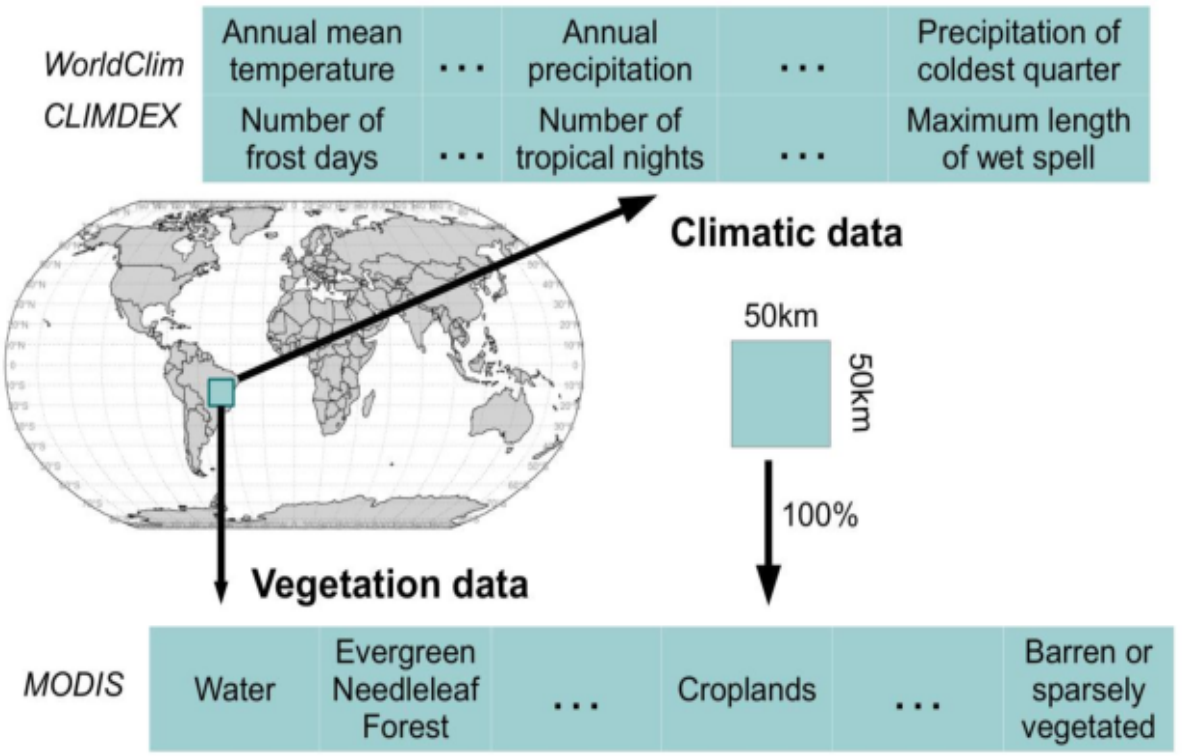


- Water
- Snow and ice
- Evergreen Needleleaf forest
- Evergreen Broadleaf forest
- Deciduous Needleleaf forest
- Deciduous Broadleaf forest
- Mixed forest
- Closed Shrublands
- Open Shrublands
- Woody Savanas
- Savanas
- Grasslands
- Permanent wetlands
- Barren or sparsely vegetated
- Urban
- Croplands
- Croplands/natural vegetation mosaic



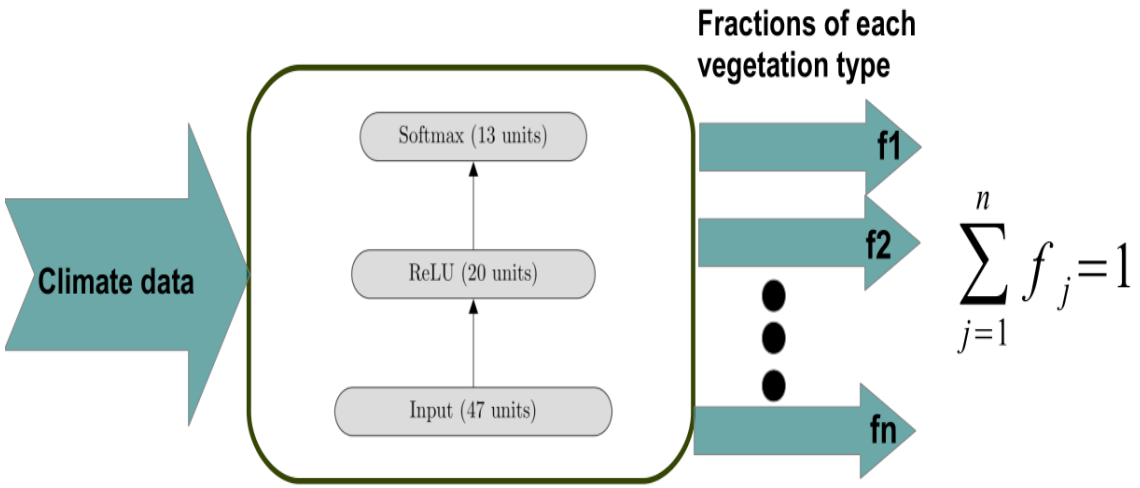
- Vegetation distribution is largely determined by climatic conditions
- As climate is changing, changes in the vegetation are expected to happen
- We are searching for patterns of how vegetation is related with climate
- Our **aim** is to predict fractions of natural landcover types from climatic conditions

Data and Challenges



- We are analyzing remote sensing data
- Predictions for each vegetation type have to be made in relation to other types
- Due to human altered landscape, most observations become incomplete

Our Approach



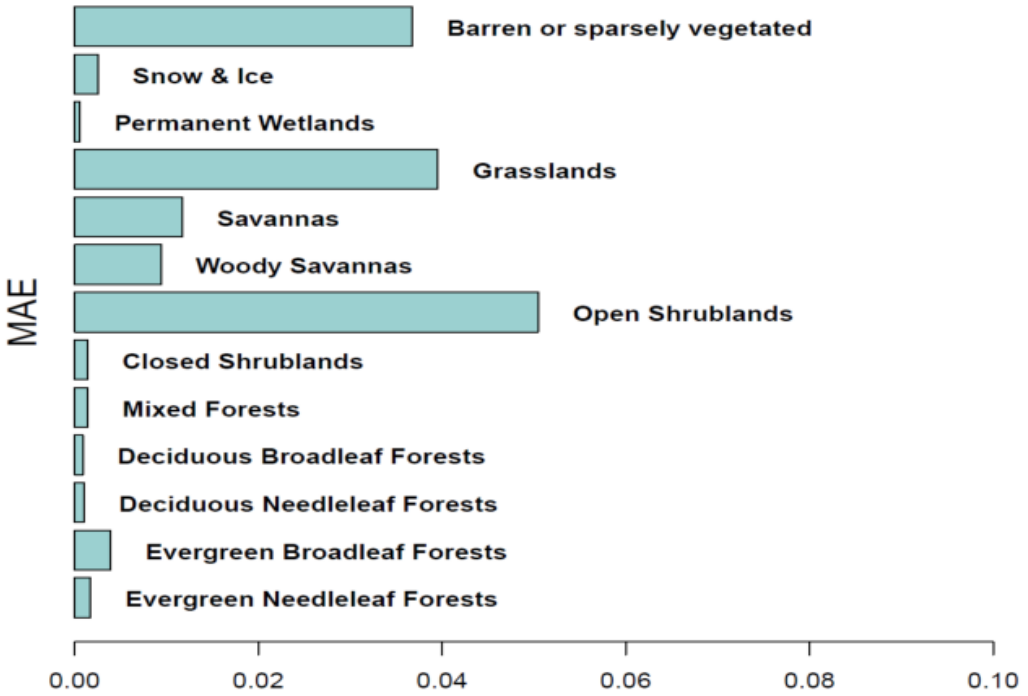
$$\sum_{j=1}^{n} f_j = 1$$

- Training the neural networks and testing the errors on complete observations
- Testing accuracy of dominant vegetation type for incomplete observations
- Experimenting on different approaches for incomplete data

While keeping the architecture of the neural network fixed, we tested six approaches for incomplete data:

1. Discarding incomplete observations
2. Re-scaling each observation to sum up to unity
3. Incomplete parts imputation based on latitudes
4. Incomplete parts imputation based on latitudes and elevation
5. Using incomplete data without alterations together with complete observations
6. Using asymmetric loss function when training neural networks

Results and Conclusions



	Complete observations	Incomplete observations
Discarding incomplete observations	94%	63%
Asymmetric loss function approach	93%	80%

Accuracy

- High average prediction accuracy can be achieved despite incompleteness of training data
- Not all vegetation types are equally well predicted
- Most accurate results are achieved using asymmetric loss function or imputing incomplete data based on latitude and elevation
- When we include incomplete data into the training set, the model is not adapted only to complete observations. In this way, we can model potentially more accurate worldwide links between vegetation and climate