



Utilizing digital twins for crop wild relatives: advancing food security and conservation efforts

24 October 2023

Desalegn Chala – Natural History Museum, University of Oslo

Nordic CWR stakeholder workshop 2023, Helsinki

- 🔥 Brief introduction to Digital Twin

- 🔥 Definition of a digital twin

- 🔥 drawing examples from different fields including life science

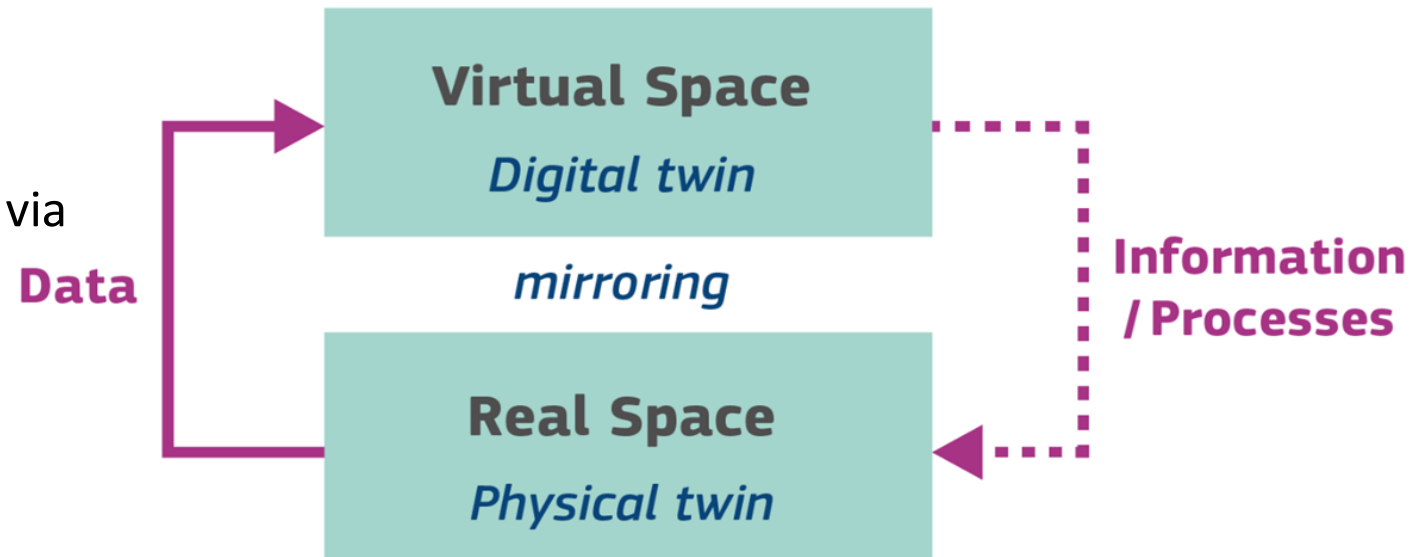
- 🔥 Brief overview of the BioDT Project

- 🔥 Crop Wild Relatives use case

- 🔥 Applications of digital twins for enhancing food security and conservations of CWR

💡 A digital counter part of a physical object or system in real time/near real time












- 💡 DT a live digital replica of an object or a process
- 💡 It is connected to the physical object or process via data streams



[Image: digital-strategy.ec.europa.eu](https://digital-strategy.ec.europa.eu)

- 🔥 Meteorologists simulate the earth's atmosphere to predict and provide weather information in a real time
- 🔥 Weather information for any location on earth is available through smartphones
- 🔥 The weather is a physical system and what we access through our smart phones is its digital counter part or digital twin

I dag 19. oktober

Tid	Vær	Temp.	Nedbør mm	Vind(kast) m/s	Vindbeskrivelse
13		3°		5 (9) ✓	Lett bris fra nord med vindkast på 9 m/s
14		3°		5 (10) ✓	Lett bris fra nordøst med vindkast på 10 m/s
15		3°		6 (10) ✓	Laber bris fra nordøst med vindkast på 10 m/s
16		2°		6 (10) ✓	Laber bris fra nord med vindkast på 10 m/s
17		3°		7 (11) ✓	Laber bris fra nordøst med vindkast på 11 m/s
18		2°		8 (13) ✓	Frisk bris fra nordøst med vindkast på 13 m/s
19		2°		8 (14) ✓	Frisk bris fra nordøst med vindkast på 14 m/s
20		2°		7 (14) ✓	Laber bris fra nordøst med vindkast på 14 m/s
21		2°		7 (12) ✓	Laber bris fra nordøst med vindkast på 12 m/s
22		2°		7 (12) ✓	Laber bris fra nordøst med vindkast på 12 m/s
23		2°		8 (13) ✓	Frisk bris fra nordøst med vindkast på 13 m/s

Weather in Finland

[Weather now](#)

Tomorrow

Day after tomorrow

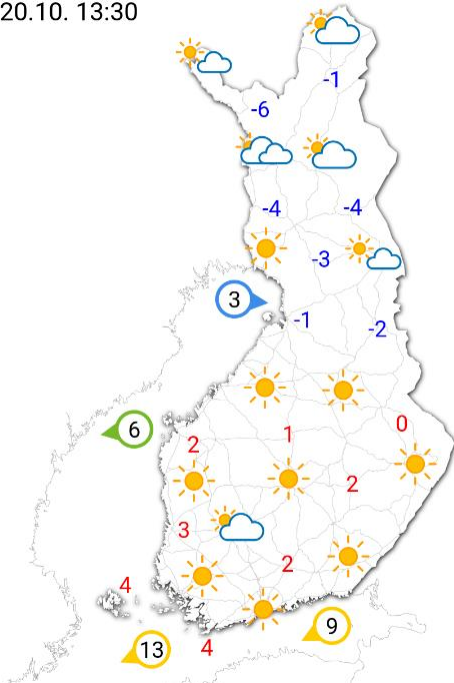
Max temperature

Min temperature

5 day precipitation sum

Snow depth

20.10. 13:30



Fire alarm systems for example

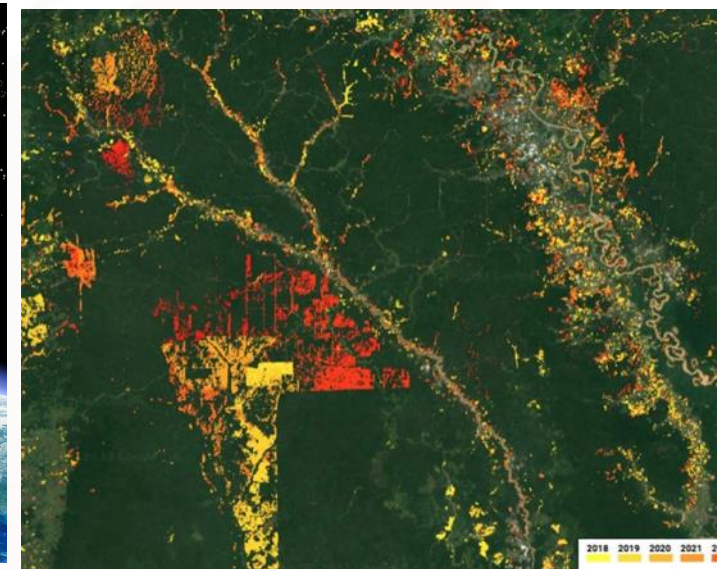
- ❖ Many homes are equipped with sensors that can detect smoke and trigger alarms when the smoke concentration is above certain threshold
- ❖ The sensors are often monitored by a central alarm company, which can notify the fire department in the event of an alarm activation
- ❖ The smoke is a physical system, and it is digitally streamed to a central server providing information in a real time



- Using satellite-based remote sensing, nowadays it is possible to monitor forest disturbances in near real time
- Disturbances due illegal logging, forest fire etc. can be detected in near time and stakeholders are automatically alerted



Satellite	Spatial Resolution	Revisiting Time	Owning Company	Number of Bands	Coverage Area
WorldView-4	31 cm	1.1 days	DigitalGlobe	4	Global Coverage
Sentinel-2	10 m - 60 m	5 days	European Space Agency (ESA)	13	Global Coverage
Landsat 8	30 m	16 days	NASA/USGS	11	Global Coverage
Pleiades	50 cm	1.5 days	Airbus Defence and Space	4	Global Coverage
GeoEye-1	41 cm	3 days	GeoEye	4	Global Coverage
SPOT 6/7	1.5 m - 6 m	1-3 days	Airbus Defence and Space	4	Global Coverage
RapidEye	5 m	5 days	Planet	5	Global Coverage
Gaofen-1	2 m - 8 m	4 days	China National Space Administration (CNSA)	4	Global Coverage
IKONOS	82 cm	3-5 days	GeoIQ	4	Global Coverage
Cartosat-2	0.6 m - 2.5 m	5 days	Indian Space Research Organisation (ISRO)	1	Global Coverage



Land Use Change Alerts (LUCA) - CTrees

- ❖ We can automate generation of different vegetation indices from satellite images for near real time monitoring of signs of
 - ❖ Abiotic and biotic stresses as soon as they appear.
 - ❖ maturity and readiness to be harvested
- ❖ Stakeholders can be alarmed and able to virtually monitor



- It is possible to fetch molecular data of medically important viruses and bacteria to
 - Prevent disease outbreaks
 - Obtain early signs of the mutations that can lead to vaccine and drug resistances

PNAS

LETTER



Reply to Ekström and Ottersen: Real-time access to data during outbreaks is a key to avoid a local epidemic becoming a global pandemic

Nils Chr. Stenseth^{a,b,c,1}, Rudolf Schlatte^d, Xiaoli Liu^e, Roger Pielke Jr.^{b,f}, Bin Chen^g, Ottar N. Bjørnstad^{b,h}, Dimitri Kusnezovⁱ, George F. Gao^j, Christophe Fraser^{k,l}, Jason D. Whittington^{a,b}, Peng Gong^m, Dabo Guan^{n,o}, and Einar Broch Johnsen^{d,1}

Ekström and Ottersen (1) correctly observe that improved real-time data access is needed for a digital twin to be effective and reliable, and that this is a caveat for our work on using digital twins to inform decisions that could prevent local epidemics becoming global pandemics (2). Data-availability in support of operational decision-making must be an important part of a “pandemic treaty” (3). Here, we further discuss the importance of data-availability (Fig. 1).

1. As mentioned by Ekström and Ottersen (1), pathogen genomics is crucial for phylogeography: tracing the origin and movements of the pathogen world-wide (4).
2. For a novel zoonotic disease, data on pathogen spread is essential to understand its transmission dynamics [e.g., by droplets or direct contact, or via an intermediate host—like mosquito in the case of malaria (5)].
3. With real-time data access, a digital twin may leverage

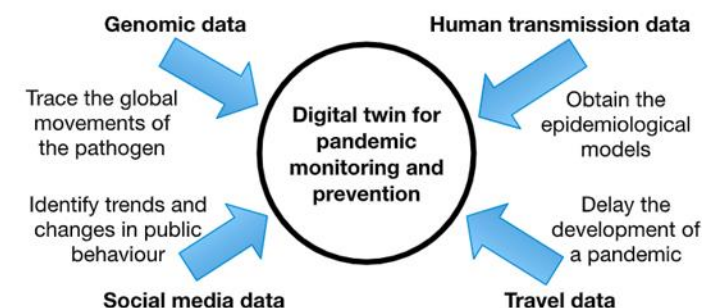


Fig. 1. Real-time access to epidemic-related data.

data rights, privacy, and aggregation, as well as agreements with nongovernmental entities that might host relevant data-sets. These issues are nontrivial and need to be addressed in a “pandemic treaty” (9).

🔥 Satellite data:

- 🔥 Multiple satellites orbit the earth, capture images with different temporal and spectral resolutions
 - 🔥 This allows us to monitor restoration sites, protected areas, afforestation programs, and real estate farms

🔥 Hyperspectral data:

- 🔥 Drones and handheld sensors provide hyperspectral images, enabling the identification of plants at species and sometimes even population levels
 - 🔥 This data has become more accessible and contributes to monitoring of invasive species and has a lot of other applications

🔥 Acoustic or sound data:

- 🔥 Publicly shared sound recordings of animals, including insects, birds, and mammals, allow us to identify species and even delve deeper.
 - 🔥 Analyzing these sounds aids in species identification and monitoring

🔥 Dynamic models:

- 🔥 Static models, such as distribution modeling, can now be automated to run at regular intervals using updated data.
 - 🔥 This dynamic approach helps to inform policymakers and conservation scientists, enabling them to make well-informed decisions

🔥 Genetics:

- 🔥 In the realm of genetics, we can create tools that automatically access genetic data from online databases, re-run phylogenetic trees, and provide alerts on changes in the Tree of Life and PhyloCode with updated information.

Horizon Europe



- 🔥 **BioDT – Biodiversity Digital Twin**
- 🔥 **A three year project**
- 🔥 **Eight Work Packages (WP) with 140+ members of different expertise backgrounds**
- 🔥 **Coordinator: CSC – IT Center for Science**
- 🔥 **Website: www.biodt.eu**

Project Information

BioDT

Grant agreement ID: 101057437



DOI

[10.3030/101057437](https://doi.org/10.3030/101057437)

Start date

1 June 2022

End date

31 May 2025

Funded under

Research infrastructures

Total cost

€ 11 059 061,00

EU contribution

€ 11 059 061,00



Coordinated by

CSC-TIETEEN TIETOTEKNIKAN KESKUS OY

+ Finland



Use Cases split into four groups

Species response to environmental change



 Biodiversity dynamics

 Ecosystem services

Genetically detected biodiversity



 Crop wild relatives and genetic resources for food security

 DNA detected biodiversity, poorly known habitats

Dynamics and threats from and for species of policy concern



 Invasive species

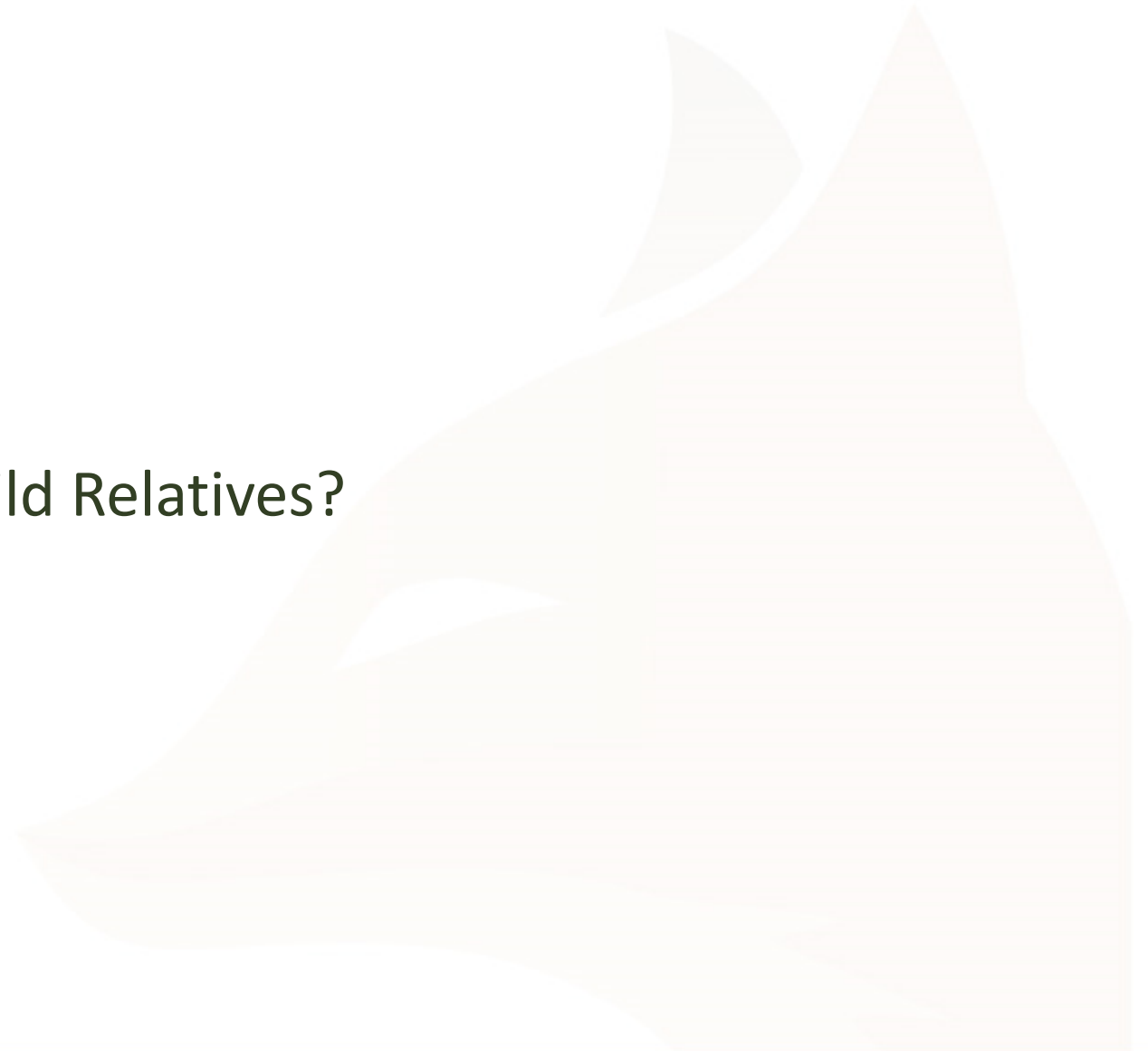
Species interactions with each other and with humans



 Disease outbreaks

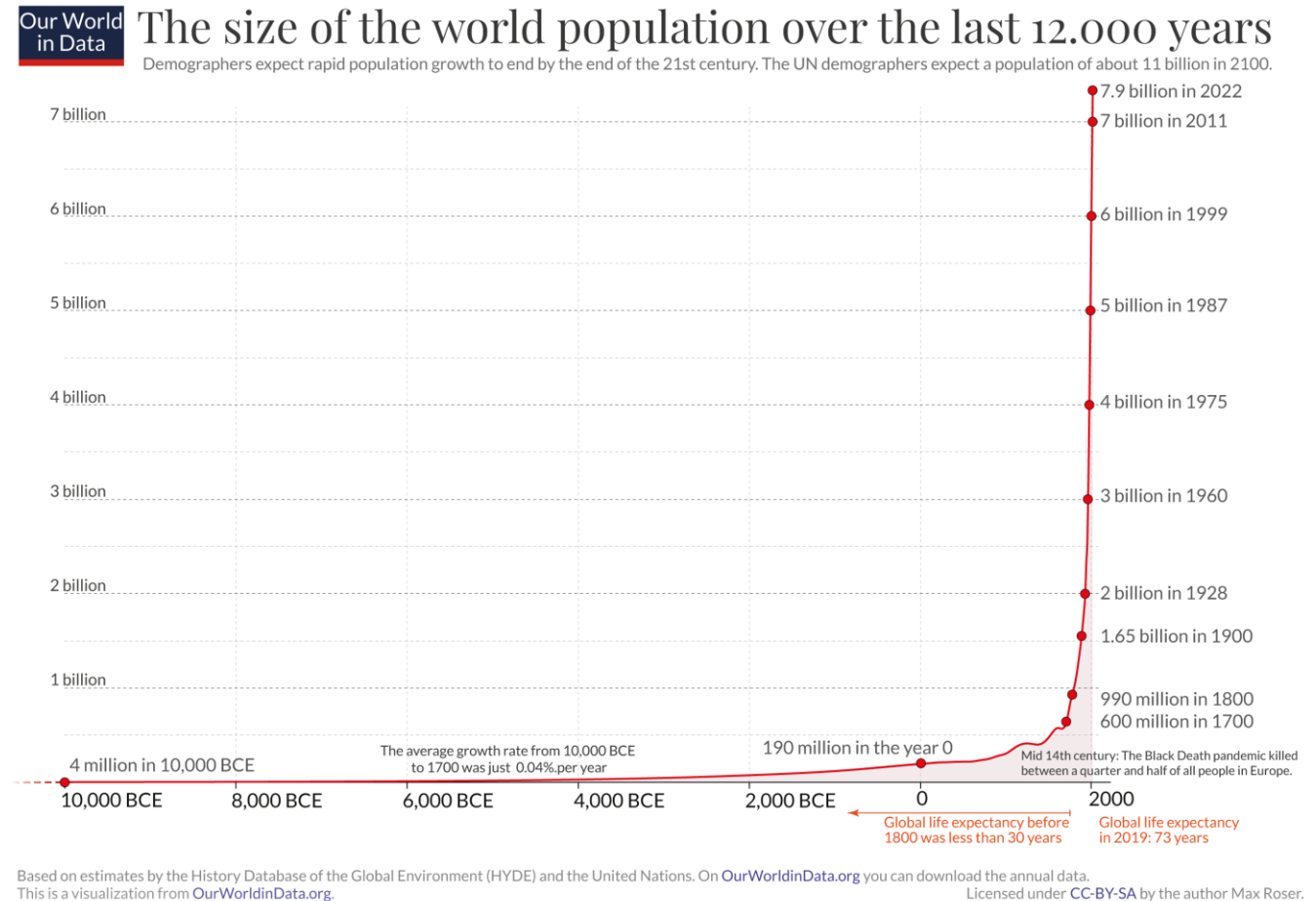
 Pollinators

Why Crop Wild Relatives?



1) Human population is growing at unprecedented rate

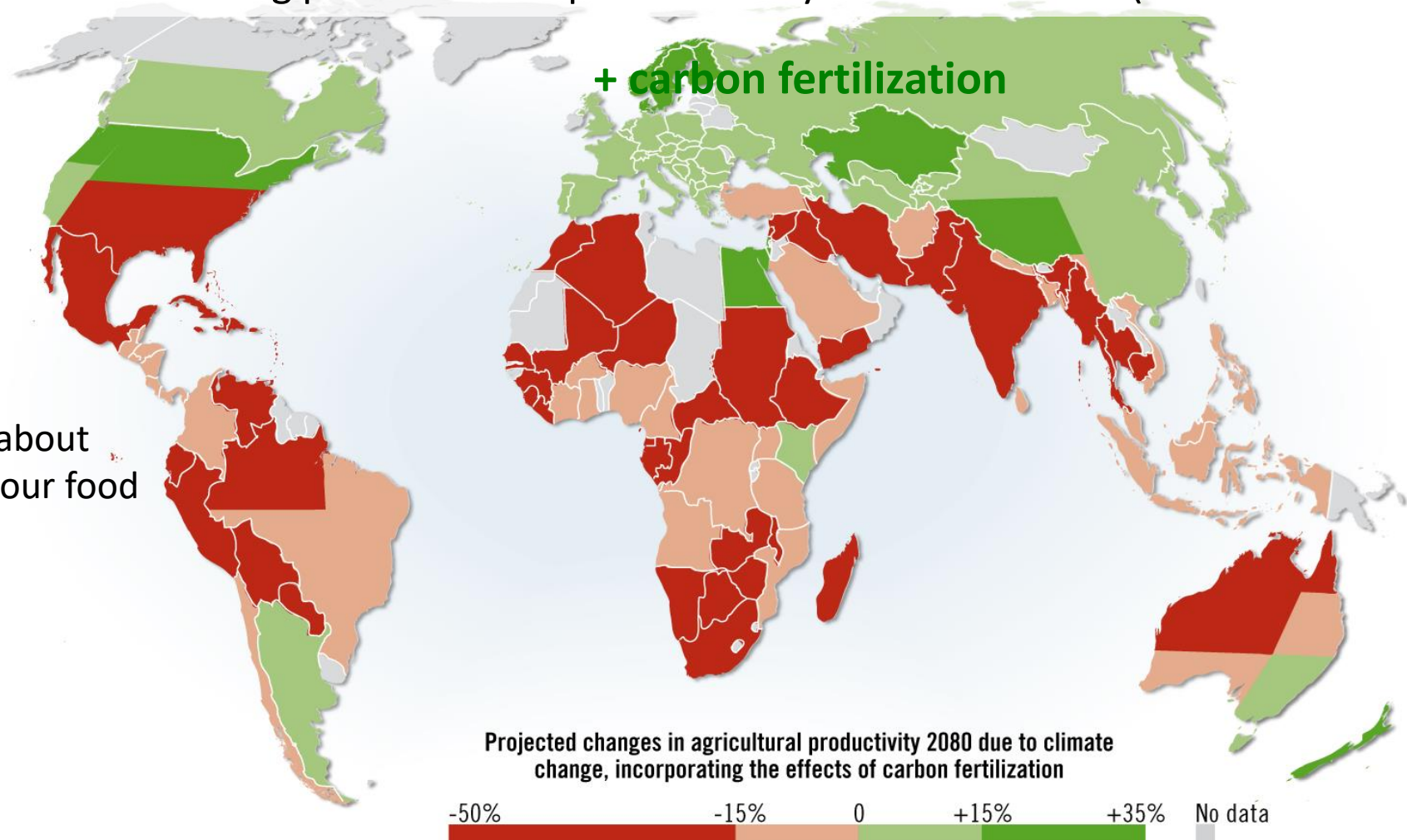
- 🔥 Since its evolution until early 19th century, the maximum number a human population achieved was one billion
- 🔥 A drastic increase over the past 200 years
 - 🔥 From one to seven billion
 - 🔥 About 700% increment
- 🔥 Expected to reach 11 billion by the end of this century



2) Potential agricultural production is challenged by global Change

Environmental stresses are reducing potential food production by 2% each decade (IPCC AR5 WGII, 2014).

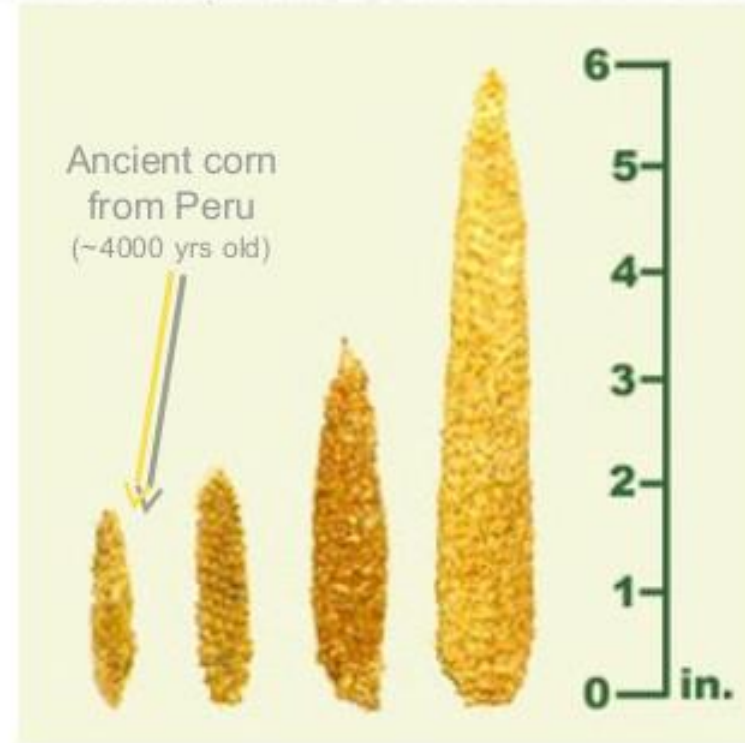
But to meet the Sustainable Development Goal of bringing about zero hunger, we need to boost our food grain production by 70%



Map by Hugo Ahlenius, GRID-Arendal (2008).Source: Cline W. (2007, 2008). Global Warming and Agriculture.

3) Domesticated gene pool is limited by human induced selection pressure

- ❖ We were selecting the traits of our interest
 - ❖ Genetic homogenization
 - ❖ Loss of genetic diversity and some adaptive traits
 - ❖ To diseases and
 - ❖ Extreme conditions
- ❖ The domesticated gene pool (genetic diversity in crops and breeding lines) is limited by the “domestication bottleneck”





🔥 Untapped genetic diversity
can be found in:

🔥 Traditional cultivars

🔥 Landraces

🔥 **Crop Wild Relatives**

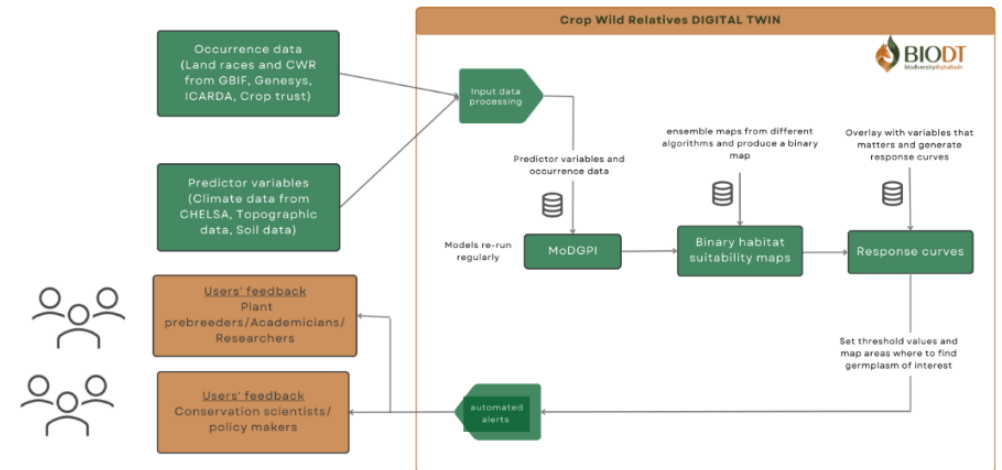
- ❖ They have been in nature since the time of their evolution
- ❖ They have been through several selection pressures against both biotic and abiotic stresses
- ❖ Through natural selection they have developed adaptive traits to:
 - ❖ Drought conditions
 - ❖ Waterlogged areas
 - ❖ Warmer temperature
 - ❖ Colder temperature
 - ❖ Saline soil
- ❖ Some are resistant to insect and fungal pests as well as other diseases
- ❖ They have important but untapped genetic resources to cope with the changing globe and to feed the ever-growing population



- 🔥 Creating a modelling tool that facilitates the search for germplasm of interest from CWR and traditional cultivars gene pool to improve domesticated crops

MoDGP: modelling the distribution of germplasm of interest

- MoDGP uses different high performing species distribution modelling algorithms
 - to produce habitat suitability maps of model targets
- Evaluate tolerances for abiotic and biotic stresses
- Rank populations of CWR based on their range of tolerances

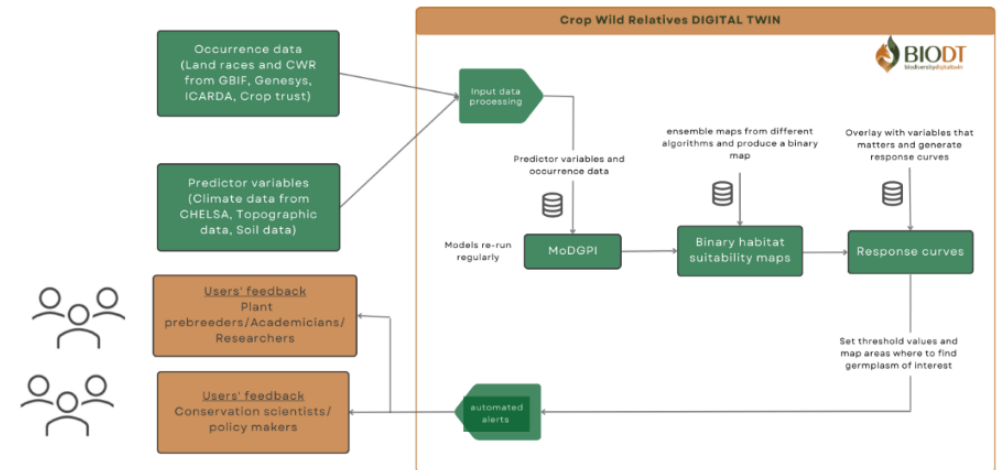


MoDGPI (Modelling the distributions of germplasm of interest)

- It uses multiple algorithms of species distribution models to produce habitat suitability of landraces and crop wild relatives
- Maps from high performing algorithms will be ensemble and classified into suitable and unsuitable classes
- Values of environmental variables that predict germplasm of interest will be extracted at each of the grid cells (pixels) that present suitable classes and pixel counts will be plotted as a response curve and threshold of values will be decided
- user interface enables users to make modifications to the threshold values and make decisions on which germplasm to test or to collect

MoDGP: modelling the distribution of germplasm of interest

- Map geographic areas where stress tolerant population possessing the desired genotypes are growing
- Automated to regularly run with updated data and provide alerts on new predicted traits of interests



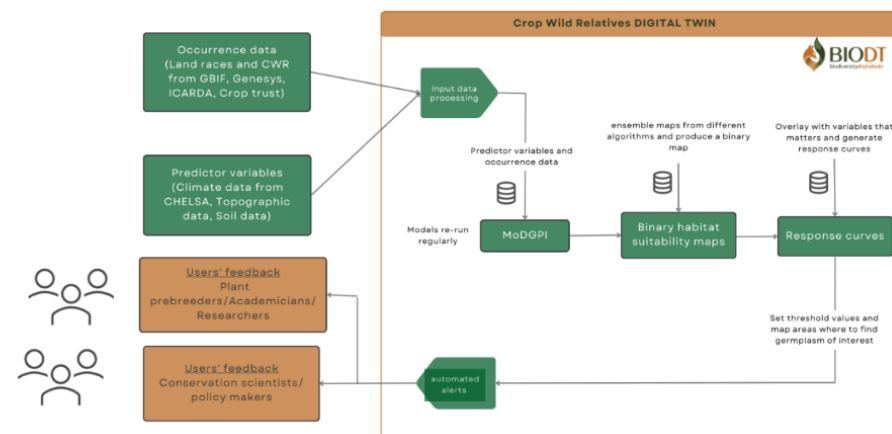
MoDGP (Modelling the distributions of germplasm of interest)

- It uses multiple algorithms of species distribution models to produce habitat suitability of landraces and crop wild relatives
- Maps from high performing algorithms will be ensemble and classified into suitable and unsuitable classes
- Values of environmental variables that predict germplasm of interest will be extracted at each of the grid cells (pixels) that present suitable classes and pixel counts will be plotted as a response curve and threshold of values will be decided
- user interface enables users to make modifications to the threshold values and make decisions on which germplasm to test or to collect

MoDGP: modelling the distribution of germplasms of interest

Input:

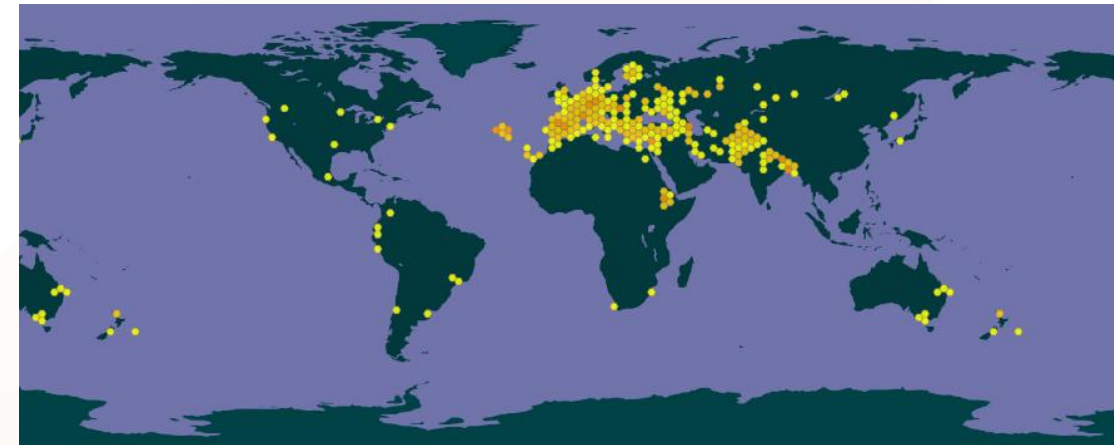
- Species occurrence data and
- Environmental variables as such as climate, soil and topographic data



MoDGPi (Modelling the distributions of germplasms of interest)

- It uses multiple algorithms of species distribution models to produce habitat suitability of landraces and crop wild relatives
- Maps from high performing algorithms will be ensembled and classified into suitable and unsuitable classes
- Values of environmental variables that predict germplasm of interest will be extracted at each of the grid cells (pixels) that present suitable classes and pixel counts will be plotted as a response curve and threshold of values will be decided
- user interface enables users to make modifications to the threshold values and make decisions on which germplasm to test or to collect

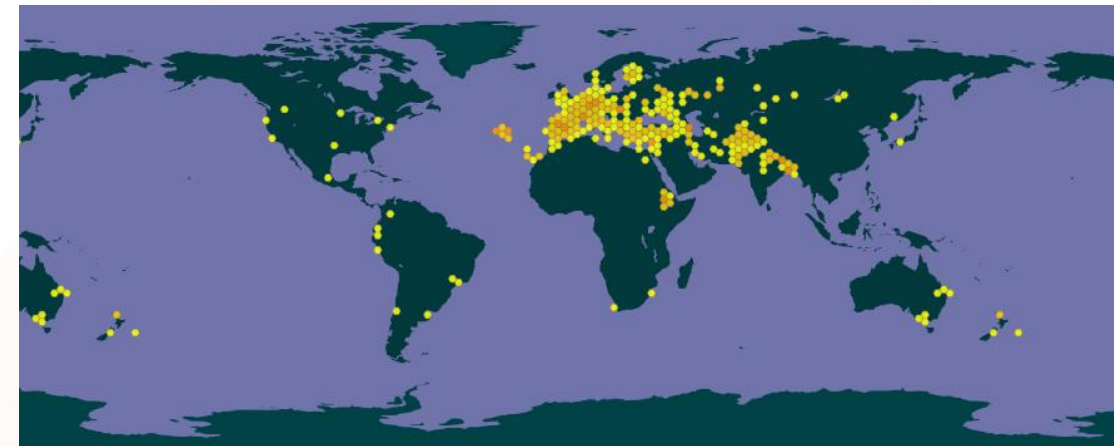
- 🔥 Belongs to the family Fabaceae
- 🔥 It can withstand extreme environments from drought to flooding
- 🔥 Cropped after the main cropping season
- 🔥 If crops fail, the same farmlands can be covered by it
 - 🔥 This makes it a climate smart species
- 🔥 Thus, grasspea is often the only alternative to starvation when other crops fail



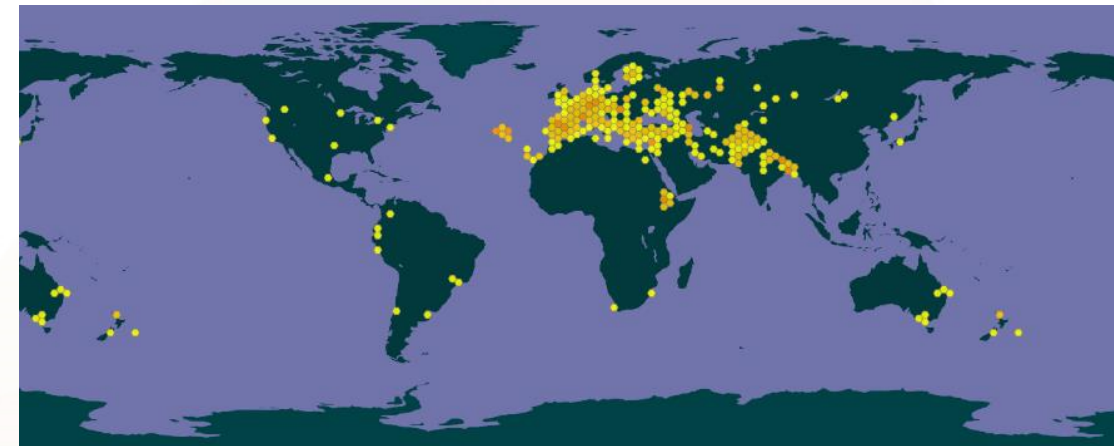
GBIF

Global Biodiversity
Information Facility

- 💧 It fixes nitrogen and serve as a natural nitrogen fertilizer
 - 💧 Can grow on degraded lands
- 💧 It also serves as a fodder
- 💧 It is very rich in protein
- 💧 Considered as a super crop to beat protein malnutrition in the future



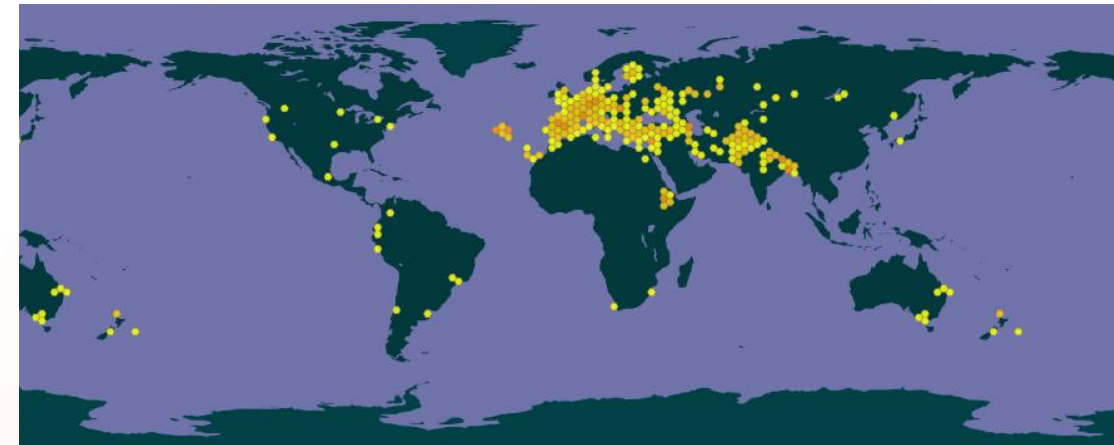
- ❖ presents a fascinating paradox – it is not only a life saver but also a destroyer as well
- ❖ It has a neurotoxin chemical
- ❖ when eaten as a large part of the diet over a long period, it can cause
 - ❖ Permanently paralyze of the lower limbs in adults - lathyrism
 - ❖ Brain damage in children
- ❖ This is often the case during famine periods



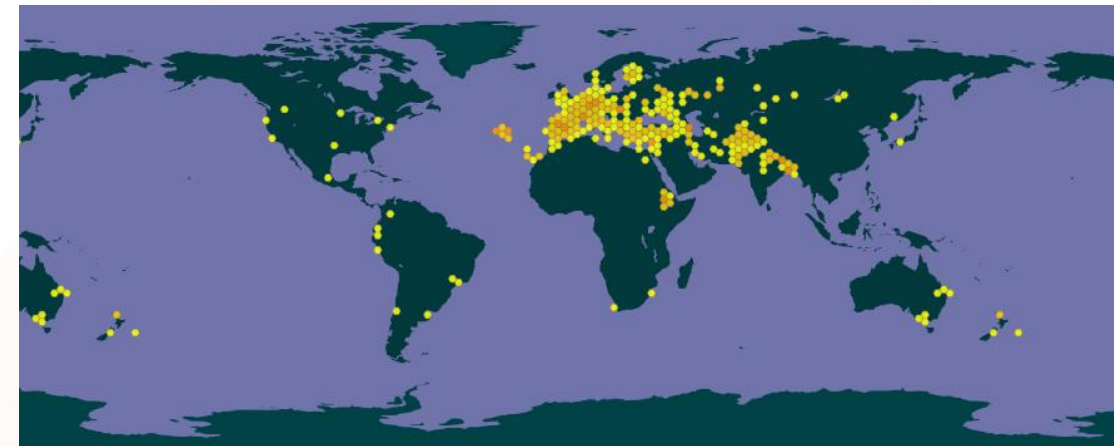
GBIF

Global Biodiversity
Information Facility

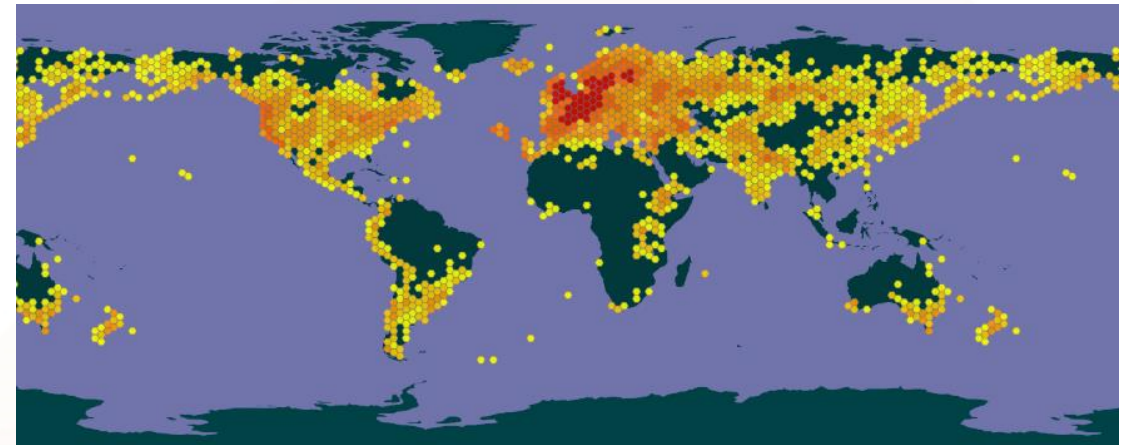
- ❖ The toxicity of grasspea is affected by different factors such as
 - ❖ water stress,
 - ❖ soil zinc content,
 - ❖ salinity
- ❖ Toxicity increases with water and zinc deficiency and with salinity.
- ❖ This means the same genotype may present different levels of toxicity under different environmental conditions, complicating the matter.



- ❖ We hypothesize that
 - ❖ 1) grasspea landraces and wild relatives growing in dry areas and in zinc deficient and acidic soils are most likely efficient in water, zinc and sodium uptake; and
 - ❖ 2) through improving these efficacies of grasspea, it is possible to minimize production of the neurotoxin.
- ❖ Thus, we are aiming to model the habitat suitability of grasspea wild relatives and identify genotypes with adaptation to these environmental factors



- ❖ Quite widely distributed
 - ❖ Diverse land races and farmers' varieties
- ❖ More than 180 herbaceous species belonging to the same genus
 - ❖ Diverse wild relatives
 - ❖ Easy to experiment with them
- ❖ Huge genetic resources
 - ❖ Certain varieties from western Asia have a low level of neurotoxin



Model targets

- 86 species out of about 180 species have reasonably good data
- with occurrence data ranging from 42 to ~200,000
- pseudo-absence/absence: 1e5 randomly selected points where the model target is absent but other species are present

```
Abs_1s <- lapply(unique(uniLoc2$species), FUN = function(x){
  NotSpecies_df <- uniLoc2[uniLoc2$species != x, ]
  SampledAbs <- sample(1:nrow(NotSpecies_df), size = 1e5, replace = FALSE)
  Report_df <- NotSpecies_df[SampledAbs,c("decimalLongitude", "decimalLatitude")]
})
```



Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
42	134	386	6853	1980	222059

🔥 Number of occurrence – 1159 unique points

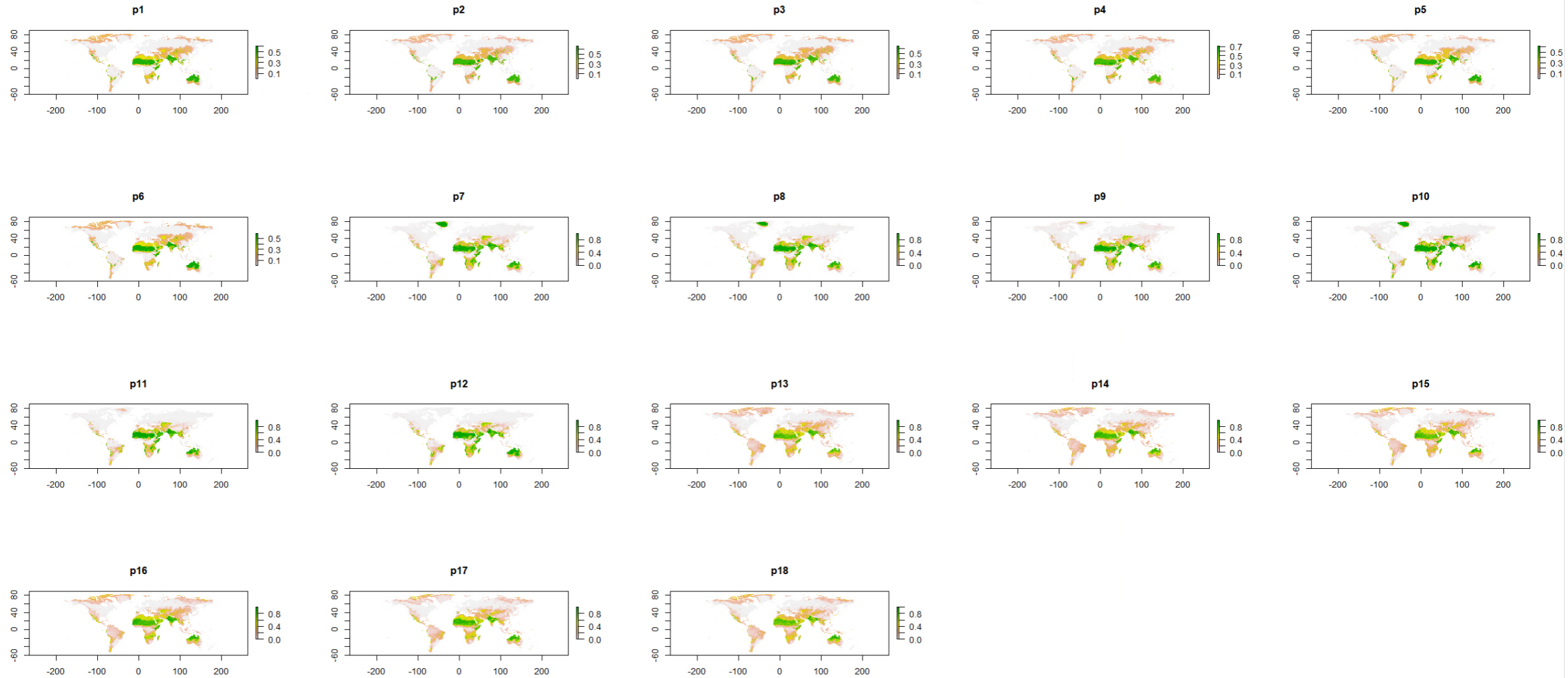
```
m = sdm(presence~., d, methods = c ("maxent","gbm","GAM","RF"),
  replications = c("sub", "boot"), test.p = 25, n = 3,
  parallelSetting = list(ncore = 4, method = "parallel"))
```

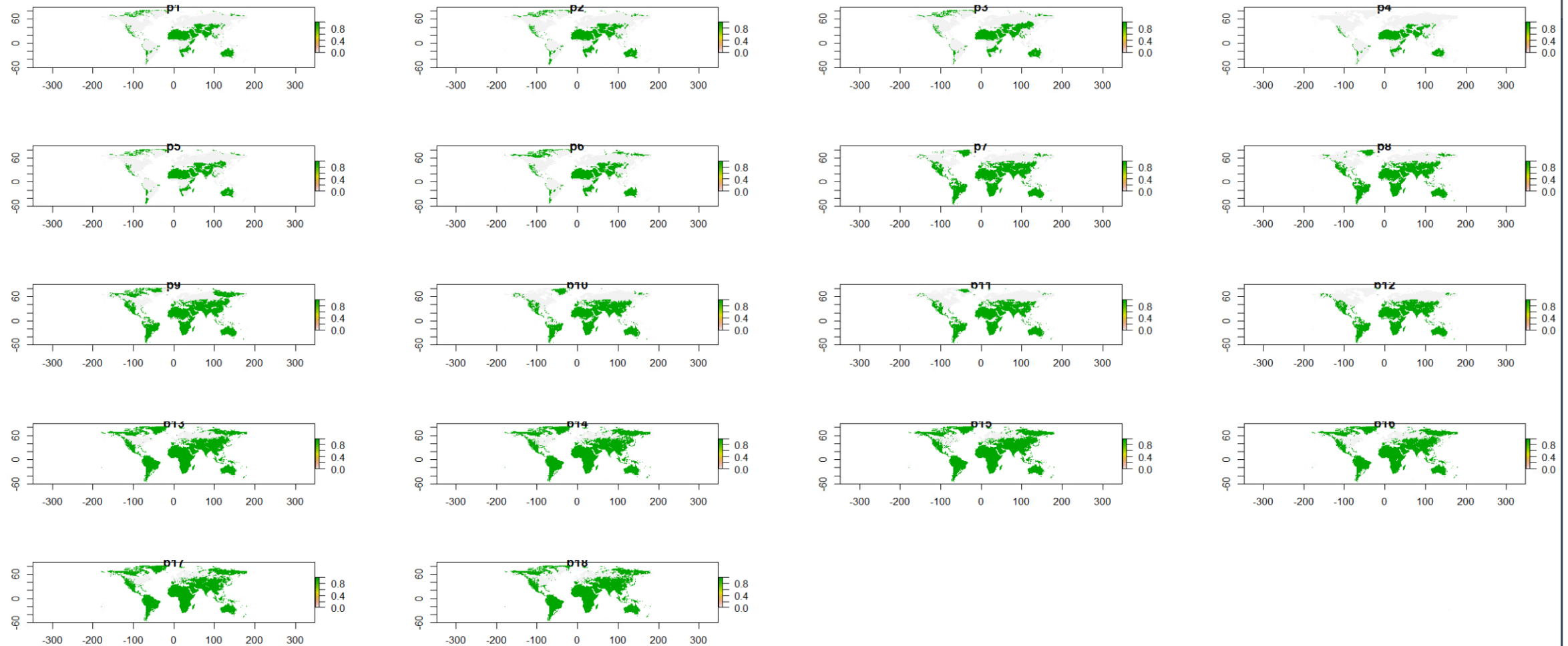
```
class                                : sdmModels
=====
number of species                    : 1
number of modelling methods          : 4
names of modelling methods           : maxent, brt, gam, rf
replicate.methods (data partitioning) : subsampling,bootstrap
number of replicates (each method)   : 3
total number of replicates per model : 6 (per species)
test percentage (in subsampling)      : 25
-----
model run success percentage (per species) :
-----
method      presence
-----
maxent      :      100 %
brt         :      100 %
gam         :      100 %
rf          :      100 %

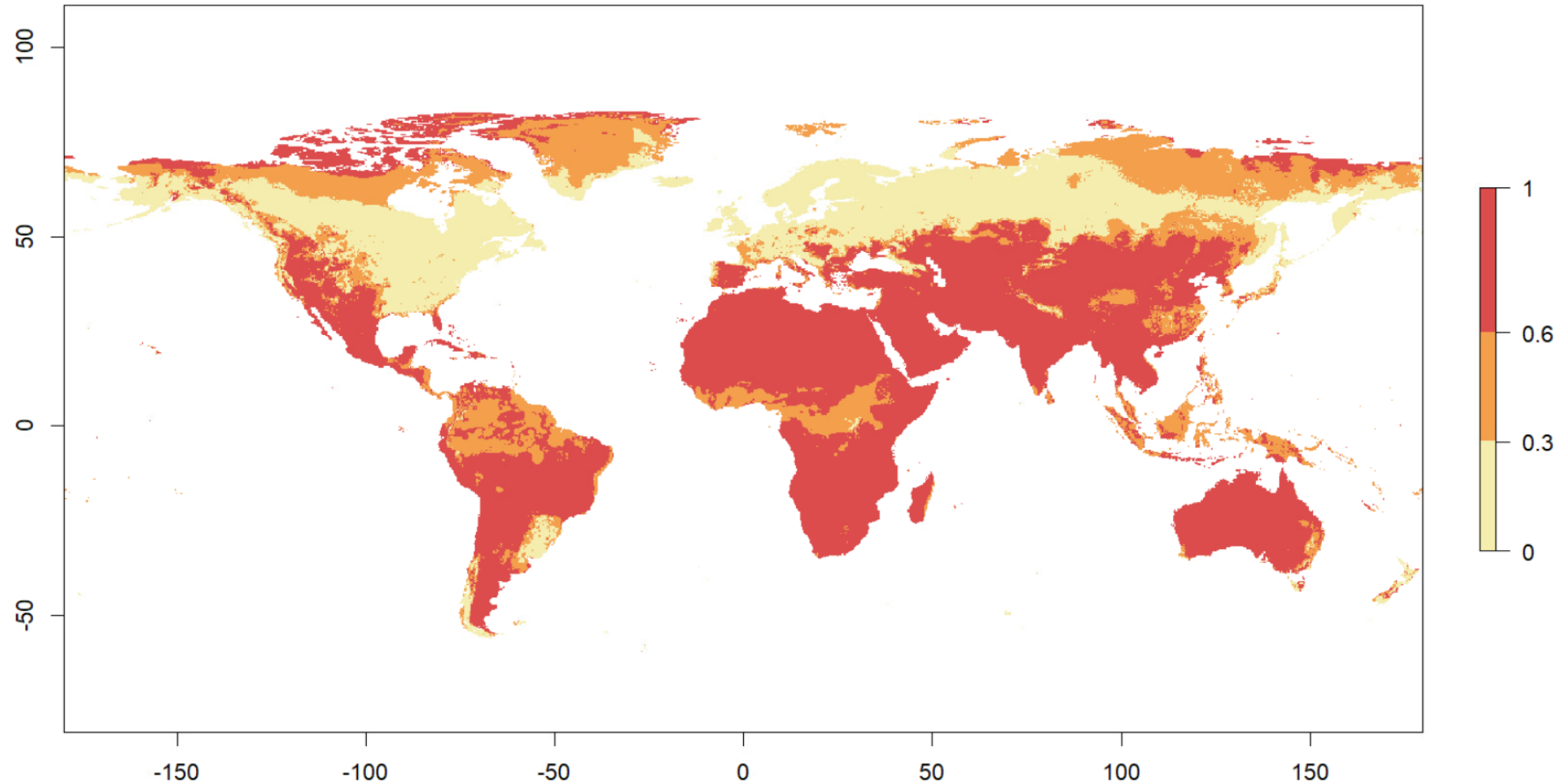
#####
model Mean performance (per species), using test dataset (generated using partitioning):
-----

## species   :  presence
=====

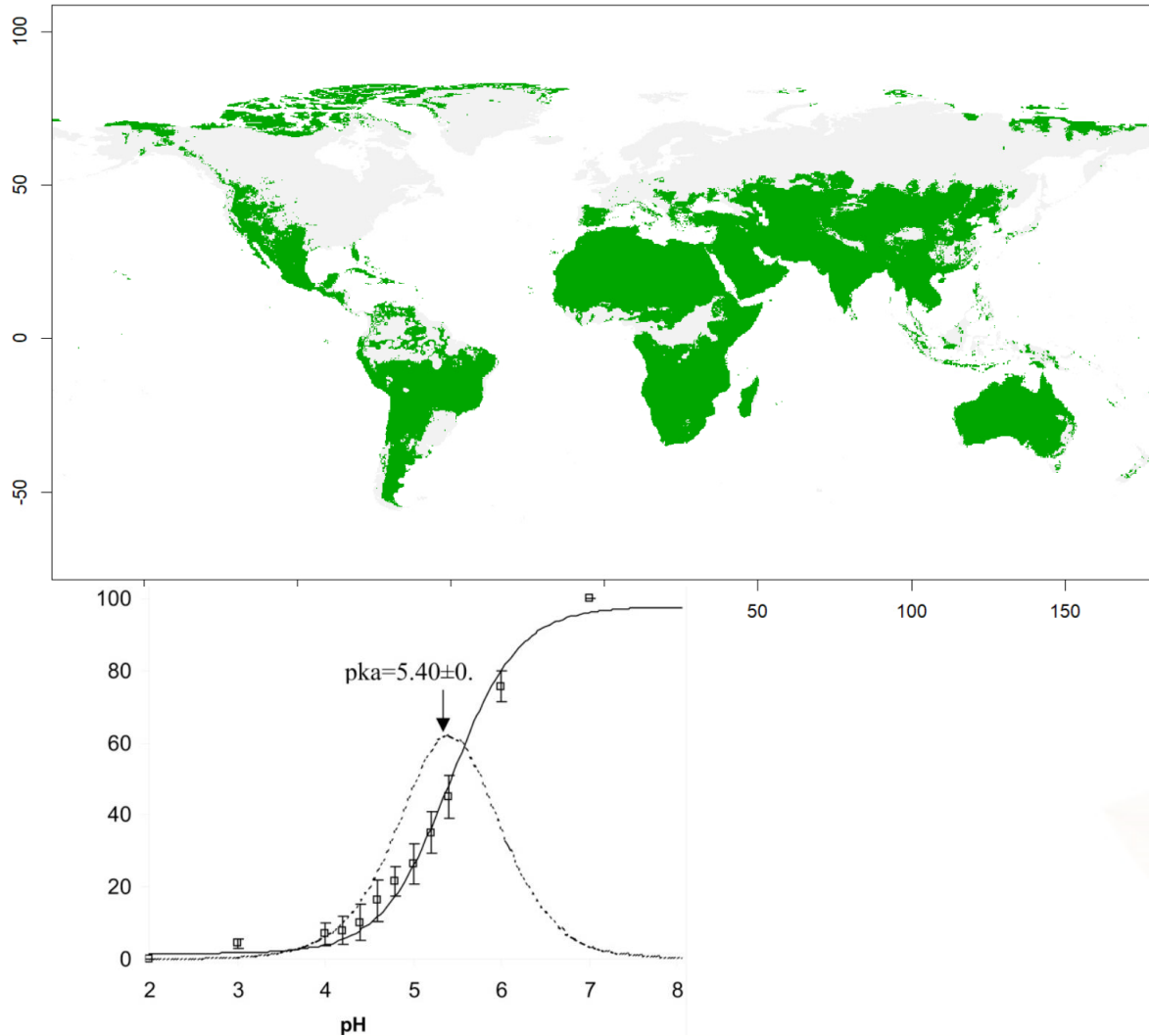
methods      :      AUC      |      COR      |      TSS      |      Deviance
-----
maxent       :      0.92      |      0.36      |      0.71      |      0.29
brt          :      0.9       |      0.48      |      0.68      |      0.09
gam          :      0.92      |      0.47      |      0.71      |      0.08
rf           :      0.96      |      0.7       |      0.81      |      0.06
```



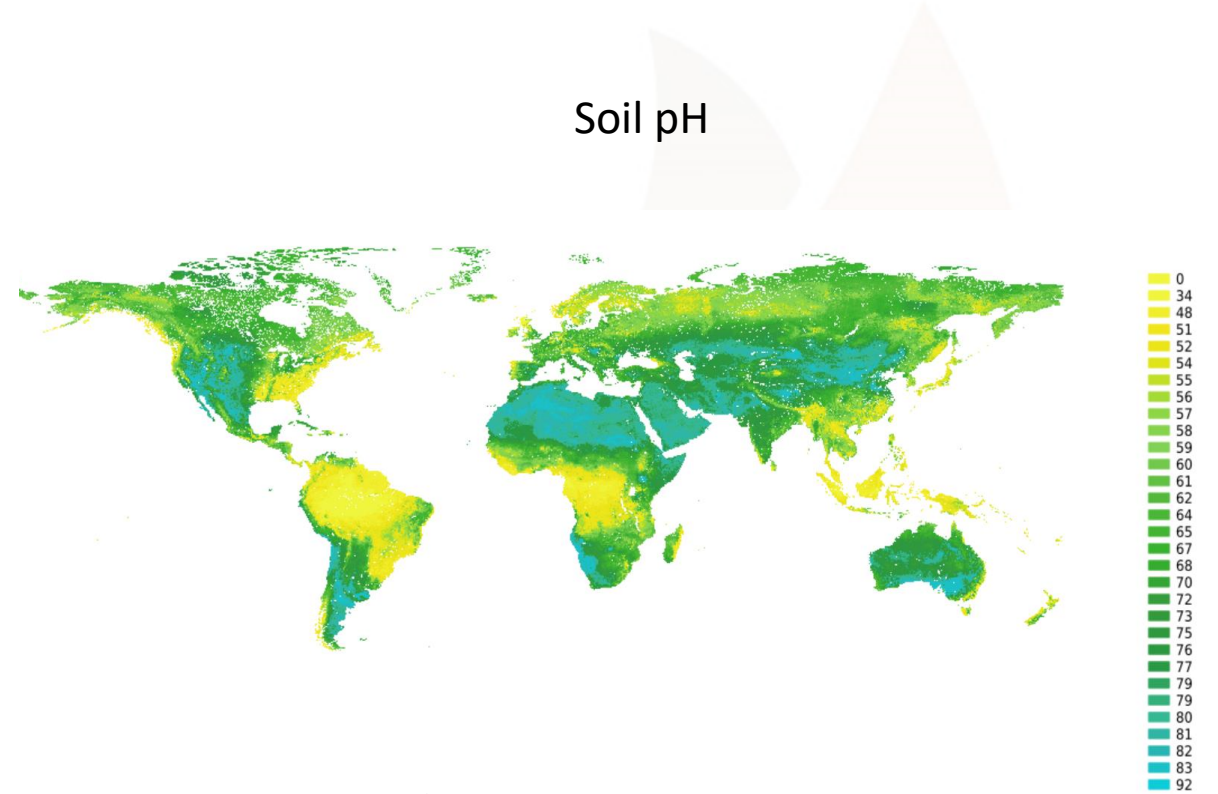




Habitat suitability



Soil pH



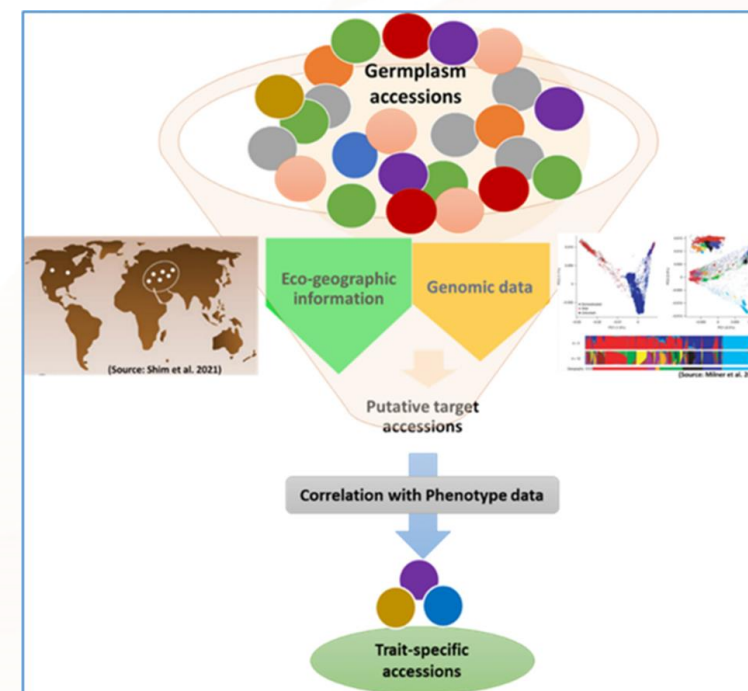
Habitat suitability for all the 86 species
 Extracting soil zinc content, soil moisture,
 pH values for suitable pixels
 Comparing ranges of tolerances
 Proposing candidate accessions to be tried (for landraces)
 Proposing candidate populations to be tried (for CWR)

Possible to provide graphical user interphase to enable users to make their alternative decisions

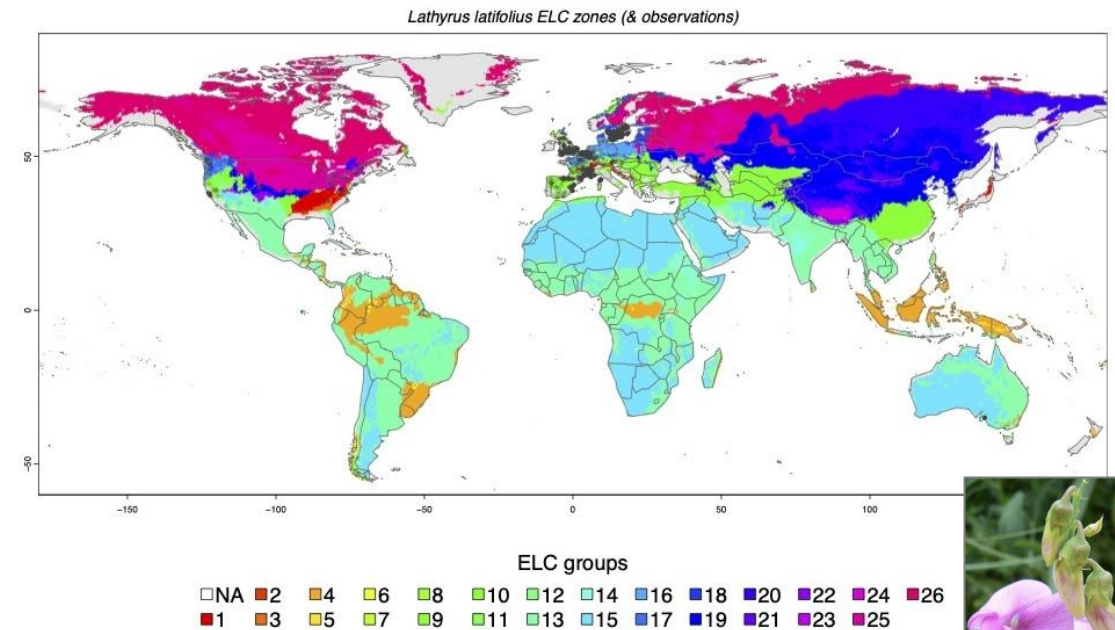
Possible to make it transferable across traits and crops

FIGS – Focused Identification of Germplasm Strategy

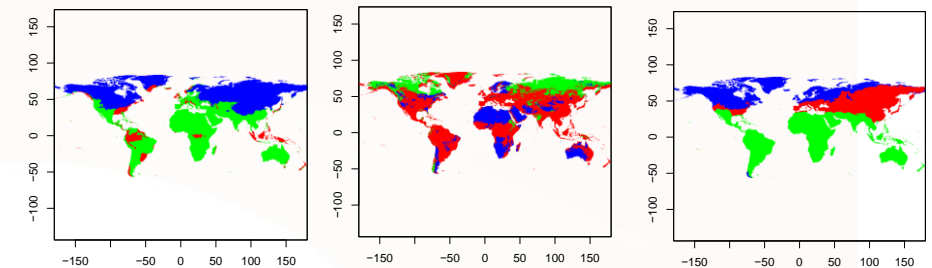
- 💧 Millions of accession in thousands of gene banks
- 💧 A lot of field trials for biotic and abiotic stress resistance
 - 💧 A lot of evaluation (trait) data (example toxicity for grasspea) from these trials
- 💧 We are linking these trait data with environmental predictors to filter accession with predicted germplasm of interest



- 🔥 Aiming at Ecotype mapping and identifying CWR population that require conservation priorities



Lathyrus latifolius



DT is used

- ✦ To automate data flow (data from distributed data sources can be fused)
- ✦ Models can be regularly updated (by feeding in updated data, followed by new model iteration)
 - ✦ Static models such as SDM can be made dynamic
- ✦ This will enhance the number of model targets and the robustness of the models over time
- ✦ Creates an opportunity to provide automated alerts for new data with predicted traits (desired alleles)



Planned for end of January 2024 at Seville , Spain (22nd to 24th)

- ❖ Participants who are good in coding, can assist in enhancing the tools we use, with a particular focus on improving data standards, FAIRification, model transferability, and developing automatic alerts for newly acquired data featuring predicted traits of interest
- ❖ There is also an option to bring one's own data and apply these models to identify populations that possess traits of interest, for purposes such as breeding or conservation
- ❖ Those who are interested in enhancing specific aspects of certain crop and need guidance on locating relevant germplasms, the hackathon can be a perfect platform



BioDT - Biodiversity Digital Twin for
Advanced Modelling, Simulation and
Prediction Capabilities project
Project number - 101057437



Horizon Europe (HORIZON)

Thank you!

