**Prediction Competition File Information**

Please see: [www.maizegxeprediction2022.org](http://www.maizegxeprediction2022.org) for competition rules and details.

Files provided to the competition participants are as follows and can be downloaded from: <https://drive.google.com/drive/folders/1leYJY4bA3341S-JxjBIgmmAWMwVDHYRb>

**PLEASE NOTE:** All data is provided as is! While we have curated and lightly filtered the data for quality we have **PURPOSELY** left much of the data in raw formats. There is also extensive missing data. These are the realities of working with large agricultural datasets! Please note that maize is grown as a hybrid crop which is the result of the cross of two inbred parents. Therefore phenotypic training and testing sets display hybrid information. The provided genotypic data set also contains hybrid information derived from inbreed genotype data. **Determining which data to use or drop, how to deal with missing values, etc. are important aspects of the competition!**

Provided Training Files:

| Training Set Files | Number of: | | |
| --- | --- | --- | --- |
| Unique Environments | Unique Hybrids | Variant Sites |
| 1\_Training\_Trait\_Data\_2014\_2021.csv | 217 | 4,683 | N/A |
| 2\_Training\_Meta\_Data\_2014\_2021.csv | 217 | N/A | N/A |
| 3\_Training\_Soil\_Data\_2015\_2021.csv | 141 | N/A | N/A |
| 4\_Training\_Weather\_Data\_2014\_2021.csv | 212 | N/A | N/A |
| 5\_Genotype\_Data\_All\_Years.vcf | N/A | 4,928 | 437,214 |
| 6\_Training\_EC\_Data\_2014\_2021.csv | 165 | N/A | N/A |

Provided Testing Set Files:

| Testing Set Files | Number of: | | |
| --- | --- | --- | --- |
| Unique Environments | Unique Hybrids | Variant Sites |
| 1\_Submission\_Template\_2022.csv | 26 | 548 | N/A |
| 2\_Testing\_Meta\_Data\_2022.csv | 26 | N/A | N/A |
| 3\_Testing\_Soil\_Data\_2022.csv | 21 | N/A | N/A |
| 4\_Testing\_Weather\_Data\_2022.csv | 26 | N/A | N/A |
| Same Genotype File as above. | N/A | N/A | N/A |
| 6\_Testing\_EC\_Data\_2022.csv | 24 | N/A | N/A |

Each file contains an “Env” column and/or a “Hybrid” column that can be used as a key for combining it with the other data files. In general, the values in these columns will match across files. However, in some cases they will not match perfectly because extra data exists for one data type or another. **It is up to each participant to determine how they prefer to treat these cases.**

**Submission Template**

The file **1\_Submission\_Template\_2022.csv** only exists for the **testing** set. This file contains the environments and hybrids for which participants will submit yield predictions. Submissions must be made using the EvalAI website (<https://eval.ai/web/challenges/challenge-page/1878/overview>) between Dec 15, 2022 and January 15, 2022. Participants should add their predictions to the template under the Yield\_Mg\_ha header and keep the other columns as they are in the template. **Submissions that do not contain all of the locations, hybrids, and headers as listed in the template, and submissions with any missing values will be REJECTED by the EvalAI evaluation script.** The columns in the submission template are as follows:

**Table 1. Description of the columns in the submission template file**

| **Name** | **Description** |
| --- | --- |
| Env | Environment (combination of location of evaluation and year) that will match fields in environmental data sets |
| Hybrid | Hybrid name that will match genotype file if it has been genotyped |
| Yield\_Mg\_ha | Grain yield in Mg per ha at 15.5% grain moisture. **(Values to be predicted by participants)** |

**Trait Data**

The file **1\_Training\_Trait\_Data\_2014\_2021.csv** contains the plot and yield data for the training set. Please note that local checks (including commercial hybrids) have their trait values set to NA. Hybrids called "Local check" are often different hybrids in different environments and commercial hybrids are typically restricted to one or a few environments. The Column descriptions are as follows:

**Table 2. Description of the columns in the trait data.**

| **Name** | **Description** |
| --- | --- |
| Env | Environment (combination of location of evaluation and year) that will match fields in environmental data sets |
| Year | Year of evaluation. |
| Field-Location | G2F Field-location name. |
| Experiment | Experiment name. |
| Replicate | Large-scale field block. |
| Block | Smaller-scale field block nested within Replicate. |
| Plot | Designation of individual experimental unit. |
| Range | Designation of field range of the plot (ranges are organized perpendicular to corn rows). |
| Pass | Designation of field pass of the plot (passes are organized parallel to corn rows. Combination of range and pass form coordinate grid system describing location of each plot within the field). |
| Hybrid | Hybrid name that will match genotype file if it has been genotyped. |
| Hybrid\_orig\_name | Hybrid name as in the DOI. |
| Hybrid\_Parent1 | Hybrid’s parental inbred line 1. |
| Hybrid\_Parent2 | Hybrid ‘s parental inbred line 2. |
| Plot\_Area\_ha | Calculated plot area in hectares. |
| Date\_Planted | Date the plot was planted. (Month/Day/Year) |
| Date\_Harvested | Date the plot was harvested. (Month/Day/Year) |
| Stand\_Count\_plants | Number of plants, per plot, at harvest. |
| Pollen\_DAP\_days | Number of days after planting that 50% of plants in the plot began shedding pollen. |
| Silk\_DAP\_days | Number of days after planting that 50% of plants in the plot had visible silks. |
| Plant\_Height\_cm | Measured as the distance between the base of a plant and the ligule of the flag leaf (centimeter). |
| Ear\_Height\_cm | Measured as the distance from the ground to the primary ear bearing node (centimeter). |
| Root\_Lodging\_plants | Number of plants per plot that show root lodging. |
| Stalk\_Lodging\_plants | Number of plants per plot broken between ground level and top ear node at harvest. |
| Yield\_Mg\_ha | Grain yield in Mg per ha at 15.5% grain moisture, using plot area without alley (Mg/ha). |
| Grain\_Moisture | Water content in grain at harvest (percentage). |
| Twt\_kg\_m3 | Shelled grain test weight (kg/m3), a measure of grain density. |

**Metadata**

The files **2\_Training\_Meta\_Data\_2014\_2021.csv** and **2\_Testing\_Meta\_Data\_2022.csv** contain the basic metadata for the locations and years in the training and test sets. The Column descriptions are as follows:

**Table 3. Description of the columns in the metadata.**

| **Name** | **Description** |
| --- | --- |
| Year | Year of evaluation. |
| Env | Environment (combination of location of evaluation and year), this will match the Env field in trait training data file. |
| Experiment\_Code | Location code associated with the experiment. |
| Treatment | Agronomic management. |
| City | Town nearest to location of trial. |
| Farm | Name of research farm on which trial was planted. |
| Field | Location within farm that trial was planted. |
| Trial\_ID (Assigned by collaborator for internal reference) | Name created by cooperator for internal use. |
| Soil\_Taxonomic\_ID and horizon description, if known | Description of soil type and/or soil horizons. |
| Weather\_Station\_Serial\_Number (Last four digits, e.g. m2700s#####) | Unique number found on the weather station. |
| Weather\_Station\_Latitude (in decimal numbers NOT DMS) | The vertical global position of the weather station. |
| Weather\_Station\_Longitude (in decimal numbers NOT DMS) | The horizontal global position of the weather station. |
| Date\_weather\_station\_placed | Date the weather station was placed in the field. |
| Date\_weather\_station\_removed | Date the weather station was removed from the field. |
| Previous\_Crop | Cultivated produce grown on the field before this trial. |
| Pre-plant\_tillage\_method(s) | How soil was prepared before planting. |
| In-season\_tillage\_method(s) | How soil was handled after planting. |
| Type\_of\_planter (fluted cone; belt cone; air planter) | Machine used to plant. |
| System\_Determining\_Moisture | Machine used to measure moisture in grain. |
| Pounds\_Needed\_Soil\_Moisture | Weight of grain needed for accurate moisture measurements. |
| Latitude\_of\_Field\_Corner\_#1 (lower left) | Lower left field corner latitude coordinate. |
| Longitude\_of\_Field\_Corner\_#1 (lower left) | Lower left field corner longitude coordinate. |
| Latitude\_of\_Field\_Corner\_#2 (lower right) | Lower right field corner latitude coordinate. |
| Longitude\_of\_Field\_Corner\_#2 (lower right) | Lower right field corner longitude coordinate. |
| Latitude\_of\_Field\_Corner\_#3 (upper right) | Upper right field corner latitude coordinate. |
| Longitude\_of\_Field\_Corner\_#3 (upper right) | Upper right field corner longitude coordinate. |
| Latitude\_of\_Field\_Corner\_#4 (upper left) | Upper left field corner latitude coordinate. |
| Longitude\_of\_Field\_Corner\_#4 (upper left) | Upper left field corner longitude coordinate. |
| Cardinal\_Heading\_Pass\_1 | Cardinal heading of first pass, i.e., direction of pass 1 looking toward end of field, (in degrees measured clockwise from N=0, S=180) |
| Issue/comment\_#1 | First submission of any issues or comments about the growing season. |
| Issue/comment\_#2 | Second submission of any issues or comments about the growing season. |
| Issue/comment\_#3 | Third submission of any issues or comments about the growing season. |
| Issue/comment\_#4 | Fourth submission of any issues or comments about the growing season. |
| Issue/comment\_#5 | Fifth submission of any issues or comments about the growing season. |
| Issue/comment\_#6 | Sixth submission of any issues or comments about the growing season. |
| Irrigated\* | Was the experiment irrigated? |
| Date\_Planted\*\* | Date the plot was planted (Month/Day/Year) |
| Plot\_Area\_ha\*\* | Calculated plot area in hectares. |
| Comments | Comments provided by the collaborator. |

\*Column present only in the file **2\_Testing\_Meta\_Data\_2022.csv.** \*\*For the training set, these columns are found on a per plot basis within the **1\_Training\_Trait\_Data\_2014\_2021.csv**.

**Soil Data**

The file **3\_Training\_Soil\_Data\_2015\_2021.csv** and **3\_Testing\_Soil\_Data\_2022.csv** contain the basic soil data for the locations and years in the training and test sets. NOTE that no soil data was collected in 2014. The Column descriptions are as follows:

**Table 3. Description of the columns in the soil data.**

| **Name** | **Description** |
| --- | --- |
| Year | Year of evaluation. |
| Env | Environment (combination of location of evaluation and year). |
| LabID | Soil laboratory which analyzed the sample. |
| Date Received | Date Ward Laboratories received the sample. |
| Date Reported | Date Ward Laboratories processed the sample. |
| E Depth | Soil sample collection depth. |
| 1:1 Soil pH | Soil pH level in a mixture, by weight, one-part soil to one-part distilled H2O. |
| WDRF Buffer pH | Woodruff method for measuring total soil acidity. |
| 1:1 S Salts mmho/cm | Soluble salts concentration in soil. |
| Texture No | Particle size analysis (PSA) with mineral components smaller than 2mm. |
| Organic Matter LOI % | Percentage of organic matter in soil. |
| Nitrate-N ppm N | Available Nitrates in parts per million (ppm). |
| lbs N/A | Amount of Nitrogen in pounds per acre. |
| Potassium ppm K | Available Potassium in ppm. |
| Sulfate-S ppm S | Available sulfate in ppm. |
| Calcium ppm Ca | Available Calcium in ppm. |
| Magnesium ppm Mg | Available Magnesium in ppm. |
| Sodium ppm Na | Available Sodium in ppm. |
| CEC/Sum of Cations me/100g | Cation exchange capacity. |
| %H Sat | Percentage of Hydrogen. |
| %K Sat | Percentage of Potassium. |
| %Ca Sat | Percentage of Calcium. |
| %Mg Sat | Percentage of Magnesium. |
| %Na Sat | Percentage of Sodium. |
| Mehlich P-III ppm P | Phosphorus extraction with dilute acid fluoride. |
| % Sand | Percentage of sand composition in soil sample. |
| % Silt | Percentage of silt composition in soil sample. |
| % Clay | Percentage of clay composition in soil sample. |
| Texture | The proportion of sand, silt, and clay sized particles that make up soil. |
| BpH | Sikora method for measuring soil acidity. |
| Zinc ppm Zn | Available Zinc in ppm. |
| Iron ppm Fe | Available Iron in ppm. |
| Manganese ppm Mn | Available Manganese in ppm. |
| Copper ppm Cu | Available Copper in ppm. |
| Boron ppm B | Available Boron in ppm. |
| Comments | Comments provided by the collaborator. |

**Weather Data**

The file **4\_Training\_Weather\_Data\_2014\_2021.csv** and **4\_Testing\_Weather\_Data\_2022.csv** contain the weather data downloaded from the NASA Power website (<https://power.larc.nasa.gov/>) for the locations and years in the training and test sets. For locations where the exact GPS field coordinates are unknown an estimate was made. Further descriptions of the weather parameters can be found on the NASA Power website.

**Table 4. Weather Column Descriptions Information.**

| **Parameter** | **Units** | **Long Name/Description** |
| --- | --- | --- |
| Env | N/A | Environment (combination of location of evaluation and year) |
| Date | N/A | Date in YYYYMMDD format |
| ALLSKY\_SFC\_PAR\_TOT | W/m^2 | All Sky Surface PAR Total |
| T2MWET | C | Wet Bulb Temperature at 2 Meters |
| QV2M | g/kg | Specific Humidity at 2 Meters |
| RH2M | % | Relative Humidity at 2 Meters |
| T2M\_MAX | C | Temperature at 2 Meters Maximum |
| ALLSKY\_SFC\_SW\_DWN | MJ/m^2/day | All Sky Surface Shortwave Downward Irradiance |
| PS | kPa | Surface Pressure |
| GWETPROF | 1 | Profile Soil Moisture |
| T2MDEW | C | Dew/Frost Point at 2 Meters |
| GWETTOP | 1 | Surface Soil Wetness |
| WS2M | m/s | Wind Speed at 2 Meters |
| T2M\_MIN | C | Temperature at 2 Meters Minimum |
| T2M | C | Temperature at 2 Meters |
| GWETROOT | 1 | Root Zone Soil Wetness |
| PRECTOTCORR | mm/day | Precipitation Corrected |
| ALLSKY\_SFC\_SW\_DNI | MJ/m^2/day | All Sky Surface Shortwave Downward Direct Normal Irradiance |

**Genotype Data**

The file **5\_Genotype\_Data\_All\_Years.vcf** found in the training set directory contains all of the genotype data for both the training and testing sets. **Please note that there are errors in this dataset, and it is minimally filtered. Please perform your own quality control before using it. Some hybrids included in the training trait data set are not included in the genotype data set because they are either commercial hybrids or their genotype data did not pass quality control filters.**

Variant calls for the 2014-2023 Genomes to Fields (G2F) materials were obtained using the Practical Haplotype Graph (PHG, Bradbury et al. 2022). The PHG aligns sequencing reads against a database populated with genome assemblies’ sequences, and imputes genotypes based on the haplotypes stored in the database.

The Maize 2.1 PHG database’s haplotypes were created from 86 genome assemblies aligned using anchorwave (Song et al. 2022) to the B73 v5 assembly (Hufford et al. 2021). 35 assemblies were downloaded from MaizeGDB (Woodhouse et al.2021), one from (Yang et al.2019), 5 Stiff-stalk assemblies from (Bornowski et al.2021), and 46 were private assemblies shared by collaborators. Nodes for the graph were defined by dividing the B73 genome into genic and intergenic reference ranges based on the Zm-B73-REFERENCE-NAM-5.0\_Zm00001eb.1.gff3. Reference range end points were selected in regions where at least 23 of the NAM genomes have 10 or more conserved base pairs. Genic regions with 0.0001 or less divergence and intergenic regions with 0.001 or less divergence were collapsed to a single consensus haplotype.

Reads for each inbred were aligned against the PHG Pangenome, and haplotypes matching each read were identified. The imputed path through the graph was then used to identify variants for the positions of the high density 600 k SNP genotyping array (Unterseer et al. 2014). Array positions were uplifted to v5 coordinate equivalents through Crossmap. Positions that could not be uplifted, missing in more than 21 assemblies, or monomorphic for all 86 assemblies were filtered out, for a final dataset with 437,214 variant positions.

Inbreds represented in the 2014-2023 G2F hybrids were sequenced with different technologies during these past 10 years:

• 2014-2017 germplasm was sequenced with GBS (genotyping by sequencing) following the Elshire et al 2011 protocol, and ApeKI as restriction enzyme. Raw data is available under bioproject IDs PRJNA385022, SRP021921, SRP009896, and SRP004282.1

• 2018-2019 germplasm was skim sequenced at ~5x coverage with Nextera libraries in a HiSeq X Ten or NovaSeq 6000. Raw data available under bioproject ID PRJNA530187

• 2020-2021 germplasm was sequenced using Exome capture at WI facilities, with a few samples sequenced as controls using GBS with ApeKI

• 2022-2023 germplasm was sequenced using GBS with PstI-MspI as restriction enzymes with the NovaSeq 6000 sequencer

A detailed list of unique genotype names (first column), to sample names (second column), to technology type (third column), and public source when available (fourth column) can be found in GenoDataSources.txt. Hybrid genotypes were generated by combining information about their parent lines using the –CreateHybridGenotypesPlugin available in TASSEL 5 (Bradbury et al. 2007). The final file of G2F hybrids contains 4,928 unique genotypes and 437,214 variant sites.

**Environmental Covariate Data**

The files **6\_Training\_EC\_Data\_2014\_2021.csv** and **6\_Testing\_EC\_Data\_2022.csv** contain 765 environmental covariates derived using an unpublished APSIM crop model developed by: Aguate, Fernando; de Leon, Natalia; de los Campos, Gustavo; Holland, James; Kaeppler, Shawn; Lima, Dayane; Lopez-Cruz, Marco; Tan, Ruijuan; Thompson, Addie; Washburn, Jacob. This simplified/generic model used 200 kg/ha of NO3 fertilization at all locations and the planting densities described in the files above. Phenological periods were estimated based on averages from the training data above and are not specific to any one hybrid.

The data headers are the Env header found in related files described above followed by the environmental covariates’ names given by a combination of a variable, a phenological period, and a soil layer. Soil layers are demarcated as 1 through 10 (1 being the top layer) with each layer consisting of 20cm in the soil column for a total of 2 meters deep. For example, the “SDR\_pGerEme\_1” column will contain environmental covariates based on the Water Supply-Demand Ratio (SDR) during the phenological period from Germination to emergence (pGerEme) within the soil layer from 0 to 20cm below ground (soil layer 1). For more in detailed explanations see the tables below and <https://builds.apsim.info/api/nextgen/docs/Maize.pdf>.

Variable names:

| **Variable** | **Description** |
| --- | --- |
| SDR | Water Supply-Demand Ratio |
| TT | Thermal Time (Celcius) |
| biomass | Above ground Biomass |
| yield | Grain yield |
| Eo | Potential evapotransipration of the whole soil-plant system (mm) |
| Eos | Potential evaporation from soil surface (mm) |
| Es | Actual (realized) soil water evaporation (mm) |
| ESW | Extractable soil water relative to LL15 (mm) |
| Flow | Water moving up (mm) |
| FlowNO3 | Amount of N leaching as NO3 from each soil layer (kg /ha) |
| Flux | Flux. Water moving down (mm) |
| Infiltration | Infiltration (mm) |
| PAWmm | Plant available water SW-LL15 (mm) |
| PotInf | Potential Infiltration: rainfall less that intercepted by the canopy and residue component |
| PotRunoff | Potential runoff (mm) |
| Runoff | Runoff (mm) |
| SWmm | Soil water content (mm) |
| T | Time since start of second state evaporation (days) |
| WaterTable | Water table |
| CoverGreen | Total plant green cover from all organs |
| CoverTotal | Total plant cover from all organs |
| LAI | Leaf area index |
| AccumulatedTT | Accumulated Thermal Time (Celcius) |

Phenological periods:

| **Abbreviation** | **Stage Description** |
| --- | --- |
| pGerEme | Germination to emergence |
| pEmeEnJ | Emergence to end of juvenile |
| pEnJFlo | End of juvenile to floral initiation |
| pFloFla | Floral initiation to flag leaf |
| pFlaFlw | Flag leaf to flowering |
| pFlwStG | Flowering to start of grain fill |
| pStGEnG | Start of grain fill to end of grain fill |
| pEnGMat | End of grain fill to maturity |
| pMatHar | Maturity to harvest/ripe |

REFERENCES:

1. Bradbury et al. The Practical Haplotype Graph, a platform for storing and using pangenomes for imputation. Bioinformatics. 2022 38:3698–702 https://doi.org/10.1093/bioinformatics/btac410

2. Woodhouse et al. A pan-genomic approach to genome databases using maize as a model system. BMC Plant Biol 2021, 385 https://doi.org/10.1186/s12870-021-03173-5

3. Yang et al. Genome assembly of a tropical maize inbred line provides insights into structural variation and crop improvement. Nat Genet 2019 51 https://doi.org/10.1038/s41588-019-0427-6

4. Bornowski et al. Genomic variation within the maize stiff-stalk heterotic germplasm pool. Plant Genome. 2021, 14: e20114 https://doi.org/10.1002/tpg2.20114

5. Hufford et al. De novo assembly, annotation, and comparative analysis of 26 diverse maize genomes. Science. 2021 373:655-662 https://doi.org/10.1126/science.abg5289

6. Unterseer et al. A powerful tool for genome analysis in maize: development and evaluation of the high density 600 k SNP genotyping array. BMC Genomics. 2014 15: 823 https://doi.org/10.1186/1471-2164-15-823

7. Bradbury et al. TASSEL: software for association mapping of complex traits in diverse samples, Bioinformatics.2007 23: 2633-2635 https://doi.org/10.1093/bioinformatics/btm308