CLIM4OMICS: a geospatially comprehensive climate and multi-OMICS database for Maize phenotype predictability in the U.S. and Canada

Parisa Sarzaeim[1], Francisco Muñoz-Arriola[1,2], Diego Jarquin[3], Hasnat Aslam[4], Natalia De Leon Gatti[5]

[1]Department of Biological Systems Engineering, University of Nebraska-Lincoln, Lincoln, NE, 68583-0726 USA, Email: parisa.sarzaeim@huskers.unl.edu
[2]School of Natural Resources, University of Nebraska-Lincoln, Lincoln, NE, 68583-0996 USA, Email: fmunoz@unl.edu
[3]Agronomy Department, University of Florida, Gainesville, FL, 32611 USA, Email: jhernandezjarqui@ufl.edu
[4]School of Natural Resources, University of Nebraska-Lincoln, Lincoln, NE, 68583-0996 USA, Email: haslam2@huskers.unl.edu
[5]Department of Agronomy, University of Wisconsin-Madison, Madison, WI 53706 USA, Email: ndeleongatti@wisc.edu

Correspondence to: Francisco Muñoz-Arriola

Abstract. The performance of numerical, statistical, and data-driven diagnostic and predictive crop production modeling heavily relies on data quality for input and calibration/validation processes. This study presents a comprehensive database and the analytics used to consolidate it as a homogeneous, consistent, and multi-dimensional genotype, phenotypic, and environmental database for maize phenotype modeling, diagnostics, and prediction. The data used is obtained from the Genomes to Fields (G2F) initiative, which provides multi-year genomic (G), environmental (E), and phenotypic (P) datasets that can be used to train and test crop growth models to understand the genotype by environment (GxE) interaction phenomenon. A particular advantage of the G2F database is its diverse set of maize genotype DNA sequences (G2F-G), phenotypic measurements (G2F-P), station-based environmental time series (mainly, climatic data) observations collected during the maize growing season (G2F-E), and metadata for each field trials (G2F-M) across the U.S. the province of Ontario in Canada, and the state of Niedersachsen in Germany. The construction of this comprehensive climate and genomic database incorporates the analytics for data quality control (QC) and consistency control (CC) to consolidate the digital representation of geospatially distributed environmental and genomic data required for phenotype predictive analytics and modeling the GxE interaction. The two-phase QC-CC pre-processing algorithm also includes a module to estimate environmental uncertainties. Generally, this data pipeline collects raw files, checks their formats, corrects data structures, and identifies and cures/imputes missing data. This pipeline uses machine learning techniques to fulfill the environmental time series gaps and quantifies the uncertainty introduced by using other data sources for gaps imputation in G2F-E, discards the missing values in G2F-P, and removes rare variants in G2F-G. Finally, an integrated and enhanced multi-dimensional database is generated. The analytics for improving the G2F database and the improved database called "CLIM4OMICS" (v1.0 and v2.0) follows the FAIR principles, and all the digital resources are available at http://doi.org/10.5281/zenodo.7490246 and http://doi.org/10.5281/zenodo.8060807, respectively.

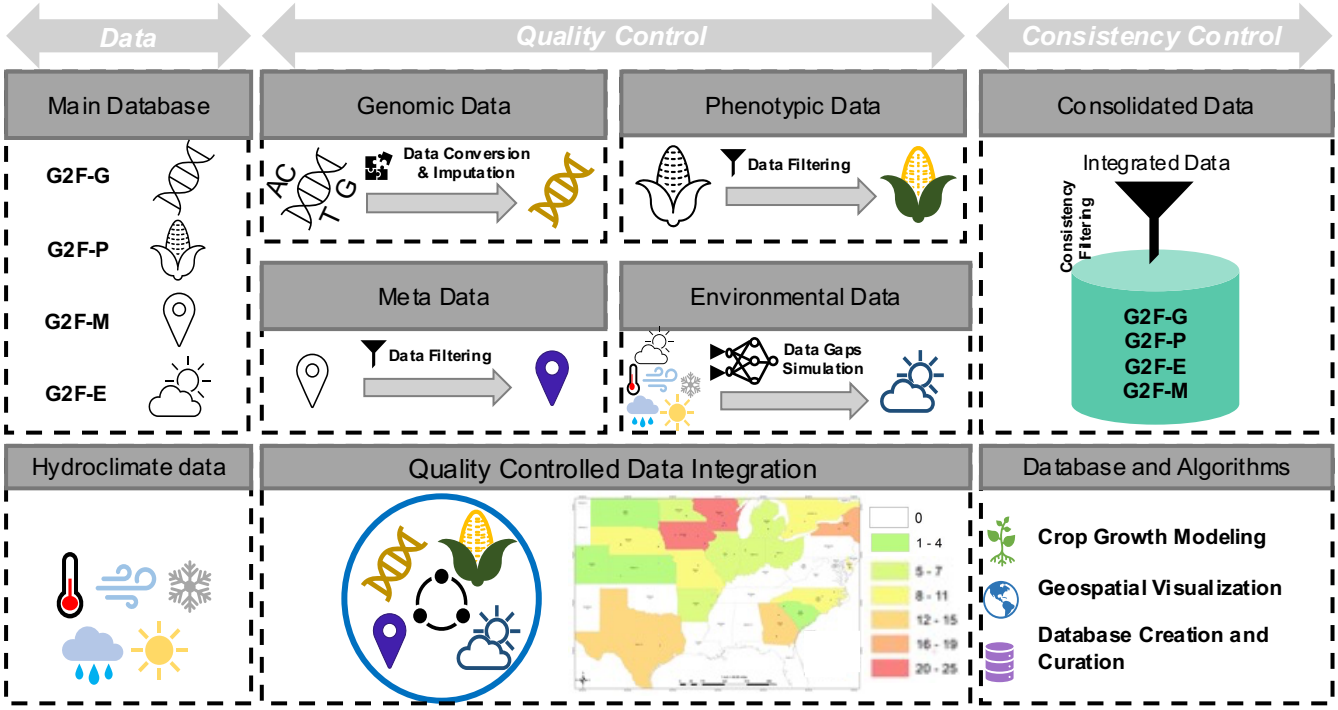


Figure 1. A conceptual framework of quality and consistency control algorithms for the multidimensional Genomes to Fields (G2F) OMICs and hydroclimatic database. "G2F-G" denotes G2F genomic data, "G2F-P" denotes G2F phenotypic data, "G2F-M" denotes G2F metadata, and "G2F-E" denotes G2F environmental data. The map indicates the locations and number of sites per state used in by the G2F initiative and represented in the CLIM4OMICS (the map is expanded as Supplementary Figure 1).

Keys: ['Genotypes', 'Positions', 'Taxa', '__DATA_TYPES__']

Genotype Lenght: ValuesViewHDF5(<HDF5 group "/Genotypes" (1579 members)>)

Shape:

<HDF5 dataset "AncestralAlleles": shape (945574,), type "<i4">
<HDF5 dataset "ChromosomeIndices": Shape (945574,), type "<i4">
<HDFS dataset "Chromosomes": shape (10,), type "|0">
<HDF5 dataset "Positions": shape (945574,), type "<i4">
<HDF5 dataset "ReferenceAlleles": shape (945574,), type "<i4">
<HDF5 dataset “SnpIds”: shape (945574,), type "|S15")

Genotypes Data:

(CML442-B
(LAMA2002-23-3-B
(LAMA2002-3S-2-B-B-B-B
(TX736)_((TX772_X_T246)_X_TX772) - 1-5-B-B-B-B-B-B6-B6-B2-B13: 100000550
(TX739)_LAMA2002-10-1-B-B-B-B3-B7_0RANGE-B: 100000510
(Tx736) ((Tx772xT246)xTx772) -1-5-B-B-B-B-B-B-B6-B12-B2-B13:100000968
(Tx739) LAMA2002- 10-1-B-B-B-B3-B7orange-B7-B11:108900969
2FACC: 100000938
2FACC: 100001100
2MCDB: 100000307
2MCDE: 100000475
3IIH6: 100000120
4N506: 100000586
511011-1-1-B: 100000114
511815-1-1-B: 100000115
511828-1-1-B: 100000142
511837-1-1-B: 100000136
511842-1-1-B: 100000119
511865-1-1-B: 100000117

Figure 2. A screenshot of the raw G2F-G data file stored in a single HDF file showing a portion of the complex hierarchical data structure of SNPs sequences.

40

Table 1. Overview of raw G2F-G data illustrating the genotyping by sequencing the molecular marker sequences of different hybrids stored in a single HDF-format file. The first column shows the maize hybrid codes and the first row shows the locus information. The A, T, G, C, and R letters are a sample of the major and minor alleles at different marker positions. The letter N denotes the missing markers in a genetic sequence at each molecular site. The source file directory for the genetic data is in "File Upload/Genotype/Markers.txt" in the database package.

	S5_6909629	S5_6909636	S5_6909641	S5_6909643	S5_6913083	S5_6913100	S5_6913110	S5_6913228	S5_6913526	S5_6913532	S5_6913539	S5_6913547	S5_6913566
BLANK:100000001	N	N	N	N	N	N	N	N	N	N	N	N	N
BLANK:100000002	N	N	N	N	N	N	N	N	N	N	N	N	N
BLANK:100000003	N	N	N	N	N	N	N	N	N	N	N	N	N
PHN11_Oh43_0075:100…	C	C	G	C	T	G	T	G	G	A	T	G	A
W10004_0248_1000000…	C	T	G	C	T	G	T	R	A	G	T	G	G
AS6103:100000006	C	T	G	C	T	G	T	G	G	A	T	G	A
PHN11_LH145_0029:10…	C	C	G	C	T	G	T	G	G	A	T	G	A
W10005_0107:1000000…	C	C	G	C	T	G	T	G	G	A	T	G	A
W10005_0032:1000000…	C	C	G	C	T	G	T	G	G	A	T	G	A
W10004_0082:1000000…	N	N	N	N	N	N	N	N	G	A	T	G	A
PHN11_LH145_0028:10....	C	C	G	C	T	G	T	G	G	A	T	G	A

Table 2. Overview of the raw G2F-P data stored in ".csv" file format showing detailed information of the phenotypic observations in 2014 as one example of the multi-year data. The "Year" column shows the year of the a specific G2F experiment, "Field-Location" column shows the 4-character name of G2F experiment consisting of the state abbreviation in the two first characters and the name of the hybrid experiment in the last two characters tested in that state, the "Recid" column shows the ID of the phenotypic record, the "Source" column shows the source of the collected phenotypic sample portal, the "Plant Height [cm]" column shows the height of the plant in [cm], the "Ear height [cm]" column shows the height of the ear in [cm], the "Stand Count [plants]" column shows the number of plants per plot at harvest, the "Root Lodging [plants]" column shows the number of plants that show the root lodging per plot, the "Stalk Lodging [plants]"column shows the number of broken plants per plot at harvest, and the "Grain Moisture [%]" column shows the percentage of the water content in plant at harvest. The other phenotypic variables have been measured and stored in similar columns. The blank cells represent the missing values of phenotypic observations. The source file directory for the phenotypic data example is in "File Upload/Phenotype/g2f_2014_hybrid_data_clean.csv" in the database package.

Year	Field-Location	RecId	Source	Pedigree	Plant Height [cm]	Ear Height [cm]	Stand Count [plants]	Root Lodging [plants]	Stalk Lodging [plants]	Grain Moisture [%]	Test Weight [lbs bu^{-1}]	Plot Weight [lbs]	Grain Yield [bu A^{-1}]
2014	DEH1	2209111	WE13-195ISO-049-X-POL-195	MOG_PHG83-129-1-1-1-1-B/LH195	186	104	40	0	4	18	54.1	10.04	98.29
2014	DEH1	2209430	13WJWE:LH185:2073	M0039/LH185	172	85	37	0	0	19.5		18.8	180.69
2014	DEH1	2209118	WE13-195ISO-390-X-POL-195.3	MOG_MO45-055-1-1-1-1-B/LH195	230	109	37	0	1	16.7	54.3	12.59	125.21
2014	DEH1	2209199	13WJWE:LH185:2865	Z022E0130/LH185	237	103	36	0	0	18.7	54.3	8.26	80.17
2014	DEH1	2209513	13WJWE:LH185:2601	W10004_0032/LH185	166	77	35	0	0	18		15.8	154.68
2014	DEH1	2209203	13WJWE:LH185:2847	Z022E0046/LH185	266	136	35	0	0	19.3	55.7	15.39	148.28
2014	DEH1	2209208	13WJWE:LH185:2661	Z013E0028/LH185	228	115	35	0	0	18.8	55.9	12.6	122.15
2014	DEH1	2209182	13WJWE:LH185:2856	Z022E0009/LH185	234	103	33	0	0	20.2	52.3	10.81	102.99
2014	DEH1	2209086	WE13-195ISO-361-X-POL-195	B73_NC230-041-1-1-1-1/LH195	227	125	26	0	0	19.3	53.9	10.86	104.63
2014	DEH1	2209169	13WJWE:LH185:2013	M0355/LH185	248	123	24	0	0	18.9	55.3	8.41	81.43
2014	DEH1	2209156	13WJWE:LH185:2214	M0172/LH185									
2014	DEH1	2209168	13WJWE:LH185:2205	M0114/LH185									
2014	DEH1	2209170	13WJWE:LH185:2073	M0039/LH185									
2014	DEH1	2209160	13WJWE:LH185:2046	M0378/LH185									
2014	DEH1	2209148	13WJWE:LH185:2055	M0266/LH185									

Table 3. Overview of raw G2F-E data stored in ".csv" file format showing the environmental time series in tabular dormat for 2014 as one example of the multi-year data. The "Record Number" column shows the number of weather station records in each experiment, the "Experiment" column shows the 4-character name of G2F experiment consisting of the state abbreviation in the two first characters and the name of the hybrid experiment in the last two characters tested in that state, the "Station ID" column shows the ID of the weather station, "NWS Network" and "NWS Station" columns show the nearest NWS network and station has been used for initial QC by the G2F collaborators, the "Day [Local]", "Month [Local]", "Year [Local]", and "Day of Year [Local]" columns show the local day, month, year, and day of year of the weather record, "Daytime [UTC]" column shows the coordinated universal time, "Temperature [C]", "Dew Point [C]", "Relative Humidity [%]", "Solar Radiation [W m^2]", " Rainfall [mm]", " Wind Speed [m s^{-1}]", Wind Direction [degrees]", and "Wind Gust [m s^{-1}] column shows the hydroclimatic time series. The blank cells represent the missing values of phenotypic observations. The source file directory for the environmental data example is in "File Upload/Environment/g2f_2014_weather.csv", in the database package.

Record Number	Experiment	Station ID	NWS Network	NWS Station	Day [Local]	Month [Local]	Year [Local]	Day of Year [Local]	Time [Local]	Datetime [UTC]	Temperature [C]	Dew Point [C]	Relative Humidity [%]	Solar Radiation [W m^2]	Rainfall [mm]	Wind Speed [m s^{-1}]	Wind Direction [degrees]	Wind Gust [m s^{-1}]
191	DEH1	9079	DE_ASOS	GED	13	5	2014	133	14:00:00	5/13/14 18:00	22.89	14.33	58.2	942	0	4.47	18	7.6
192	DEH1	9079	DE_ASOS	GED	13	5	2014	133	14:30:00	5/13/14 18:30	21.78	13.89	60.5	918	0	4.92	40	7.6
193	DEH1	9079	DE_ASOS	GED	13	5	2014	133	15:00:00	5/13/14 19:00	21.56	13.17	58.4	855	0	4.02	21	6.71
194	DEH1	9079	DE_ASOS	GED	13	5	2014	133	15:30:00	5/13/14 19:30	20.83	12.89	60	778	0	4.47	14	7.15
195	DEH1	9079	DE_ASOS	GED	13	5	2014	133	16:00:00	5/13/14 20:00	20.72	12.72	59.8	728	0	4.92	351	7.15
196	DEH1	9079	DE_ASOS	GED	13	5	2014	133	16:30:00	5/13/14 20:30	20.22	12.83	62	642	0	4.02	19	6.26
197	DEH1	9079	DE_ASOS	GED	13	5	2014	133	17:00:00	5/13/14 21:00	20.06	12.67	62.1	552	0	3.58	354	5.81
198	DEH1	9079	DE_ASOS	GED	13	5	2014	133	17:30:00	5/13/14 21:30	19.28	12.89	66	452	0	4.47	5	6.26
199	DEH1	9079	DE_ASOS	GED	13	5	2014	133	18:00:00	5/13/14 22:00	17.89	12.78	71.6	350	0	4.92	32	5.81
200	DEH1	9079	DE_ASOS	GED	13	5	2014	133	18:30:00	5/13/14 22:30			75.9	284	0	4.47	25	5.81
201	DEH1	9079	DE_ASOS	GED	13	5	2014	133	19:00:00	5/13/14 23:00	16	12.5	79.6	155	0	3.58	36	5.36
202	DEH1	9079	DE_ASOS	GED	13	5	2014	133	19:30:00	5/13/14 23:30	14.94	12.22	83.7	79	0	3.58	25	6.26
203	DEH1	9079	DE_ASOS	GED	13	5	2014	133	20:00:00	5/14/14 0:00	14.06	12	87.4	8	0	4.02	33	6.26
204	DEH1	9079	DE_ASOS	GED	13	5	2014	133	20:30:00	5/14/14 0:30	13.67	12	89.8	0	0	3.13	12	5.36
205	DEH1	9079	DE_ASOS	GED	13	5	2014	133	21:00:00	5/14/14 1:00	13.22	12.17	93.3	0	0	3.13	9	5.36

Table 4. Overview of raw G2F-M data stored in ".csv" file format showing the metadata collected for the 2014 experiments as one example of the multi-year data. The "Location Name" column shows the state and the number of the experiment in that state, the "Type" column shows the type of the experiment which can be hybrid or inbred, the "Experiment" column shows the 4-character name of G2F experiment consisting of the state abbreviation in the two first characters and the name of the hybrid experiment in the last two characters tested in that state, the "City" column shows the city that the experiment located at, the "Farm" column shows the name of the farm that the experiment has been tested in, the "Field" column shows the name of the field of the experiment, and "lon" and "lat" columns show the longitude and the latitude of the weather station installed in the field. The source file directory for the metadata example is in *"File Upload/Meta/g2f_2014_field_characteristics.csv"* in the database package.

Location name	Type	Experiment	City	Farm	Field	lon	lat
DE	hybrid	DEH1	Georgetown	Elbert N. & Ann V. Carvel Research & Education Center	27AB	-75.20	38.63
GA	hybrid	GAH1	Tifton	Bellflower	18	-83.55	31.50
IA1	hybrid	IAH1	Ames	Worle		-93.69	41.99
IA2	hybrid	IAH2	Carroll			-94.72	42.06
IA3	hybrid	IAH3	Keystone			-92.25	41.98
IA4	hybrid	IAH4	Crawfordsville	Southeast Research Farm	14	-91.48	41.19
IL1	hybrid	ILH1	Urbana	Maxwell Farms	MF500	-88.23	40.06
IN	hybrid	INH1	West Lafayette	Purdue ACRE	97/98	-87.00	40.48
MN	hybrid	MNH1	Waseca	Southern Research & Outreach Center	NA	-93.53	44.06
MO1	hybrid	MOH1	Columbia	Bradford	C1a	-92.20	38.89
MO2	hybrid	MOH2	Columbia	Rollins=Hinkson Creek Bottoms	block 5	-92.35	38.92
NC	hybrid	NCH1	Kinston	Cunningham Research Farm	L block 5	-77.57	35.29
NE1	hybrid	NEH1	Lincoln	East Campus	1807	-96.65	40.83
NE2	hybrid	NEH2	North Platte	Dryland farm		-100.74	41.05
NE3	hybrid	NEH3	Brule	North Dryland	West 1/4	-101.99	41.16
NY1	hybrid	NYH1	Aurora	Musgrave Research Farm	J	-76.65	42.72
NY2	hybrid	NYH2	Aurora	Musgrave	E4	-76.65	42.73
ON1	hybrid	ONH1	Waterloo	Rosdendale	Huras	-80.42	43.49
ON2	hybrid	ONH2	Ridgetown	On Campus	Range 5	-81.88	42.45
TX1	hybrid	TXH1	College Station	University Farm	224	-96.43	30.54
TX2	hybrid	TXH2	halfway	Halfway	pivot	-101.94	34.18
WI	hybrid	WIH1	Madison	West Madison	M1400	-89.53	43.057

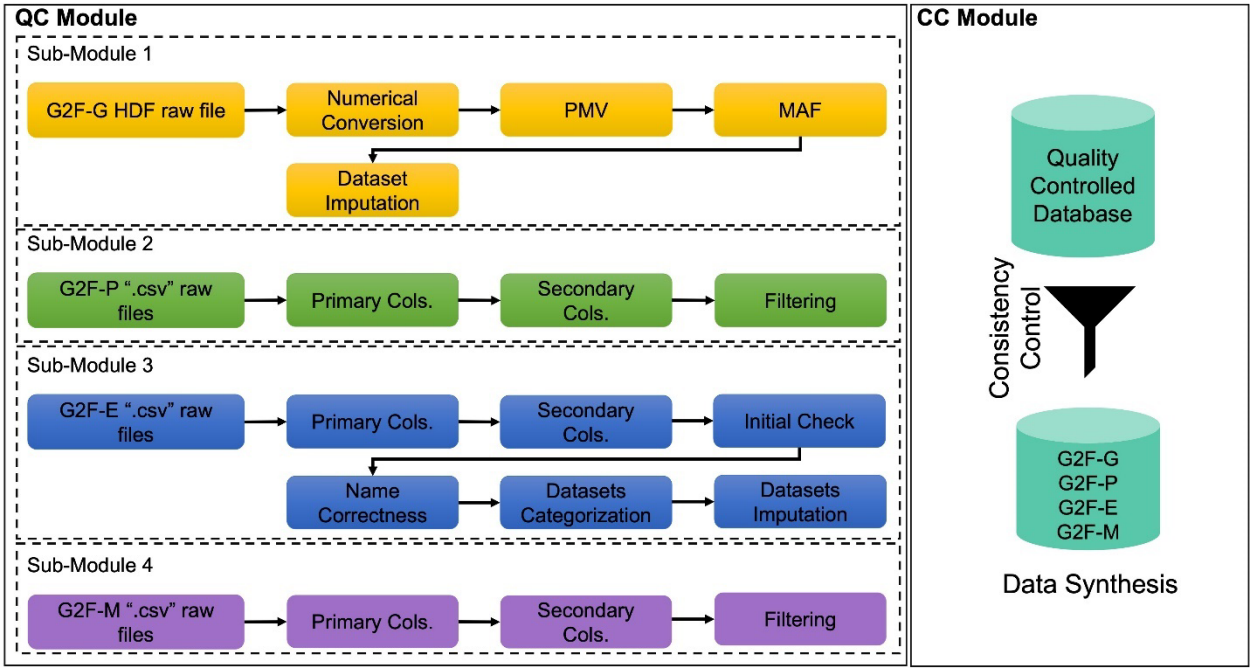


Figure 3. The overall algorithmic QC-CC framework for G2F database. The "G2F-G," "G2F-P," "G2F-E," and "G2F-M" denote the G2F genomic, phenotypic, environmental, and meta data, respectively. The "PMV" and "MAF" denote the percent missing values and minor allele frequency, respectively. The "Primary Cols." and "Secondary Cols." denote primary and secondary columns, respectively.

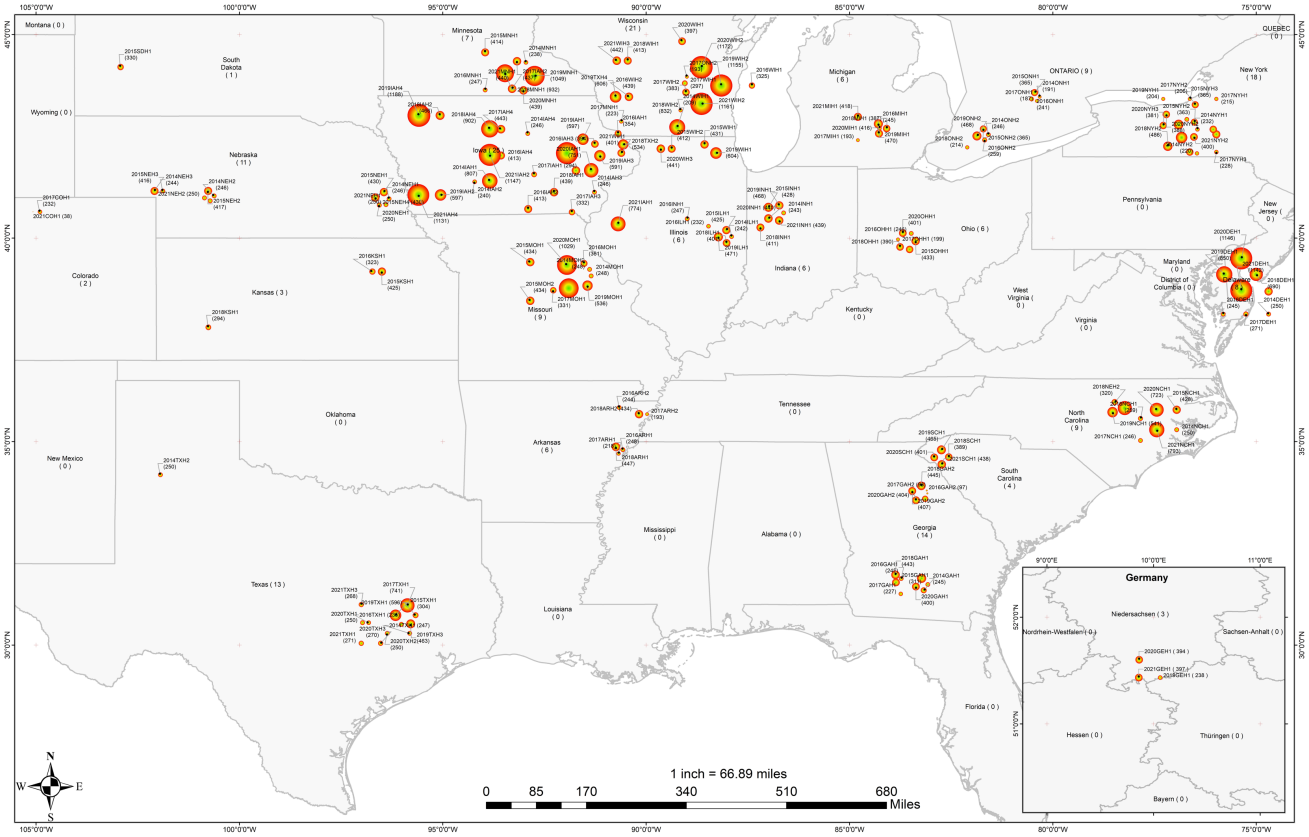


Figure 4. The spatial distribution of phenotypic records of G2F experiments in the U.S. regions and the province of Ontario in Canada between 2014. And 2021. The state of Niedersachsen in Germany includes the years 2018, 2020, and 2021 for three locations. The location of each station in the map was modified for visualization purposes, allowing the illustration of sations with multi-year records. The size of the circle represents the number of years sampled, which also appears within the parenthesis next to the year at each site. The colors of the circles were included for visualization purposes only.

90

Table 5. Record of a single of G2F-P dataset. It shows the phenotypic measurements including "Plant Height (cm)," "Ear Height (cm)," "Grain Moisture (%)," and "Grain Yield (bu A^{-1})" for a maize hybrid with pedigrees of "B37" and "MO17" collected in "2014-DEH1" experiment located in Delaware in 2014. The ID of the record is "2014_DEH1_B37/MO17, and the ID of the experiment is "2014DEH1". The "H" denotes the hybrid type of the experiment, "P1" and "P2" denote the pedigrees of the maize hybrid, and "DE" denotes the state of Delaware.

ID	Year	Location	Experiment	Experiment ID	Pedigree	P1	P2	Plant Height (cm)	Ear Height (cm)	Grain Moisture (%)	Grain Yield (bu A^{-1})
2014_DEH1 _B37/MO17	2014	DEH1	2014-DEH1	2014DEH1	B37/MO17	B37	MO17	235	139.5	19.2	217.2

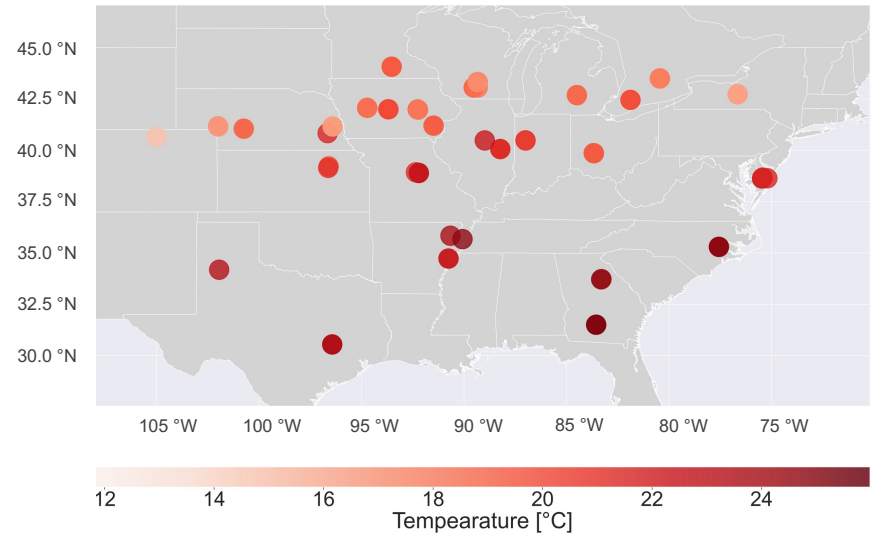
Table 6. The percentage of complete, empty, and incomplete portions of time series for each G2F hydroclimatic variable: Temperature (T), Dew Point (DP), Relative Humidity (RH), Solar Radiation (SR), Rainfall (R), Wind Speed (WS), Wind Direction (WD.

	T (°C)	DP (°C)	RH (%)	SR (W m^{-2})	R (mm)	WS (m s^{-1})	WD (degrees)	
Complete	78.6	69.6	79.2	37.6	84.3	76.4	23.6	-
Empty	0	6.1	0.5	11.8	0	1.1	1.6	-
Incomplete	21.4	24.3	20.3	50.6	16.7	22.5	74.8	-

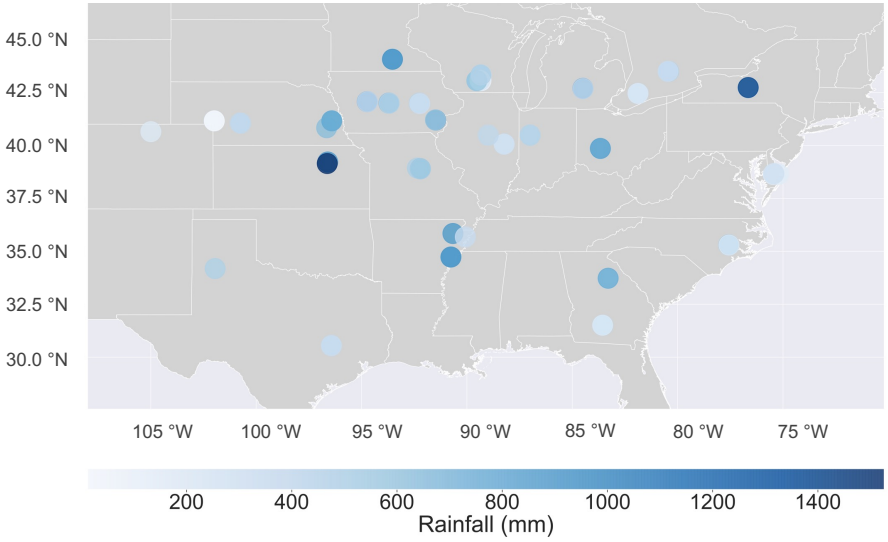
105

Table 7. Record of a single example of G2F-E dataset. It shows the observed hydroclimate data including "Temperature (°C)," "Dew Point (°C)," "Relative Humidity (%)," "Solar Radiation (W m^{-2})," "Rainfall (mm)," "Wind Speed (m s^{-1})," "Wind Direction (degrees)," and "Wind Gust (m s^{-1})" collected by weather station with ID of "9079" for "2014DEH1" experiment located in Delaware on 9 May 2014 at 15:00:00 local time. The ID of the experiment is "2014DEH1". The "H" denotes hybrid type of the experiment, and "DE" denotes the state of Delaware.

Record Number	Station ID	Location	Experiment ID	Day [Local]	Month [Local]	Year [Local]	Day of Year [Local]	Time [Local]	Temperature (°C)	Dew Point (°C)	Relative Humidity (%)	Solar Radiation (W m^{-2})	Rainfall (mm)	Wind Speed (m s^{-1})	Wind Direction (degrees)	Wind Gust (m s^{-1})
1	9079	DEH1	2014DEH1	9	5	2014	129	15:00:00	23.06	15.78	63.2	887	0	1.79	32	4.02



(a)



(b)

Figure 5. The spatial distribution of (a) improved mean temperature (T_{mean}) and (b) improved accumulated rainfall (R_{acc}) records in G2F-E database during the maize growing season in all G2F experimental fields in 2014-2017.

120

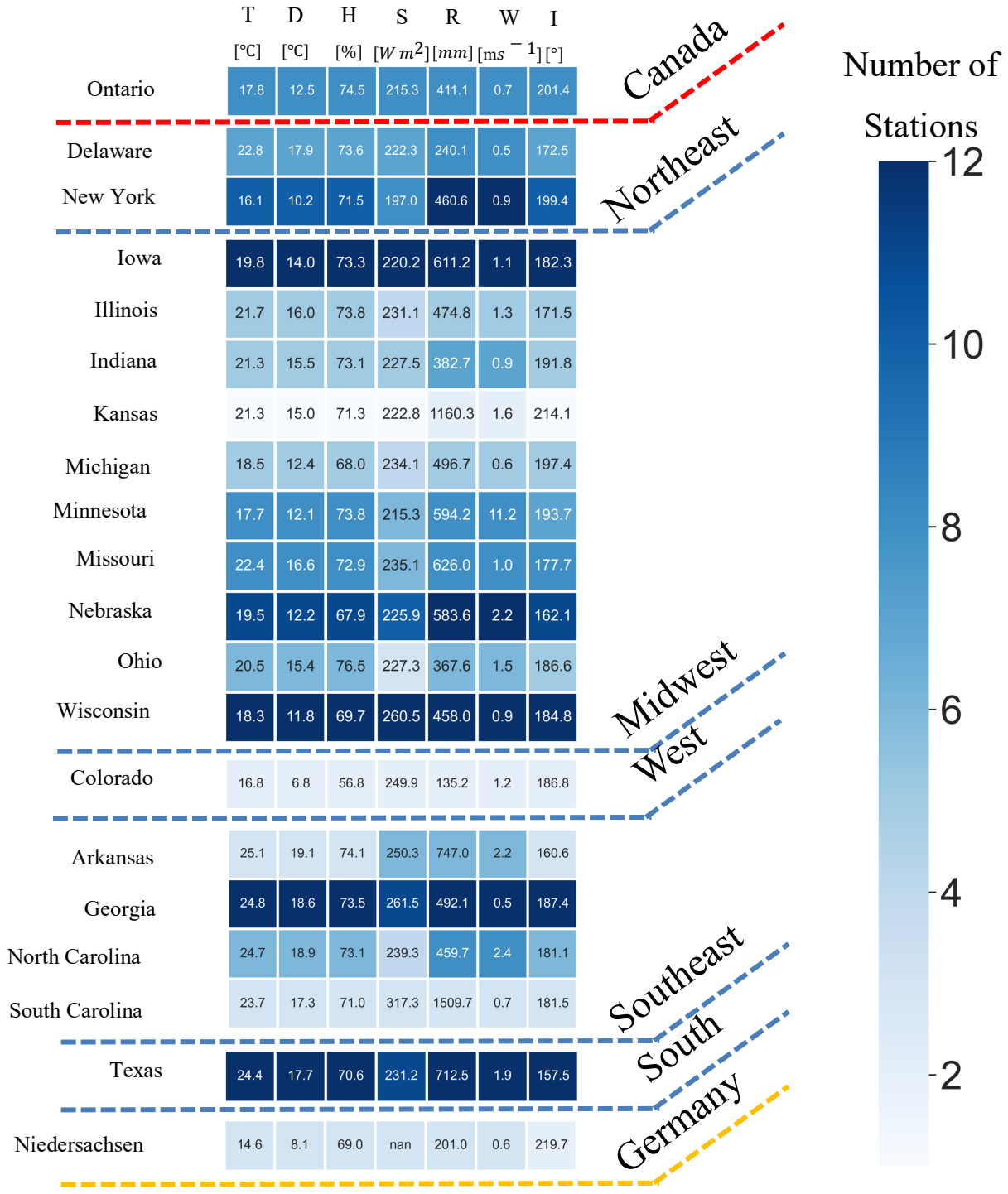

T [°C]
D [°C]
H [%]
S [W m2]
R [mm]
W [ms − 1]
I [°]
Ontario 17.8 12.5 74.5 215.3 411.1 0.7 201.4
Delaware 22.8 17.9 73.6 222.3 240.1 0.5 172.5
New York 16.1 10.2 71.5 197.0 460.6 0.9 199.4
Iowa 19.8 14.0 73.3 220.2 611.2 1.1 182.3
Illinois 21.7 16.0 73.8 231.1 474.8 1.3 171.5
Indiana 21.3 15.5 73.1 227.5 382.7 0.9 191.8
Kansas 21.3 15.0 71.3 222.8 1160.3 1.6 214.1
Michigan 18.5 12.4 68.0 234.1 496.7 0.6 197.4
Minnesota 17.7 12.1 73.8 215.3 594.2 11.2 193.7
Missouri 22.4 16.6 72.9 235.1 626.0 1.0 177.7
Nebraska 19.5 12.2 67.9 225.9 583.6 2.2 162.1
Ohio 20.5 15.4 76.5 227.3 367.6 1.5 186.6
Wisconsin 18.3 11.8 69.7 260.5 458.0 0.9 184.8
Colorado 16.8 6.8 56.8 249.9 135.2 1.2 186.8
Arkansas 25.1 19.1 74.1 250.3 747.0 2.2 160.6
Georgia 24.8 18.6 73.5 261.5 492.1 0.5 187.4
North Carolina 24.7 18.9 73.1 239.3 459.7 2.4 181.1
South Carolina 23.7 17.3 71.0 317.3 1509.7 0.7 181.5
Texas 24.4 17.7 70.6 231.2 712.5 1.9 157.5
Niedersachsen 14.6 8.1 69.0 nan 201.0 0.6 219.7
Canada
Northeast
Midwest
West
Southeast
South
Germany
Number of Stations
12
10
8
6
4
2

Figure 6. The heatmap for number of G2F experiments in the U.S. regions and the province of Ontario in Canada between 2014. And 2021. The state of Niedersachsen in Germany includes the years 2018, 2020, and 2021 for three locations. The color shows the number of stations in each state. The number in each cell represents the average of hydroclimatic variables in each state including mean of Temperature (T), mean of Dew Point (D), mean of Relative Humidity (R), mean of Solar Radiation (S), accumulative Rainfall (R), mean of Wind Speed (W), and mean of Wind Direction (I).

Table 8. Record of a single of G2F-M dataset. It shows the location including "Lat" and "Lon" of the "2014DEH1" experiment located in Delaware in 2014. The ID of the experiment is "2014DEH1". The "Lat" denotes latitude, "Lon" denotes longitude, "H" denotes the hybrid type of the experiment, and "DE" denotes the state of Delaware.

Experiment	Experiment ID	Experiment type	Year	State	Lat	Lon
DEH1	2014DEH1	H	2014	DE	38.63	-75.20

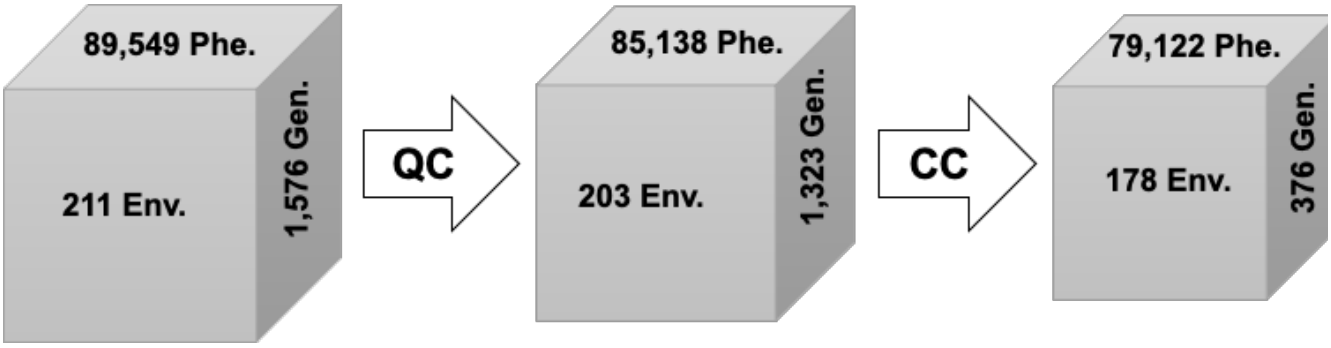
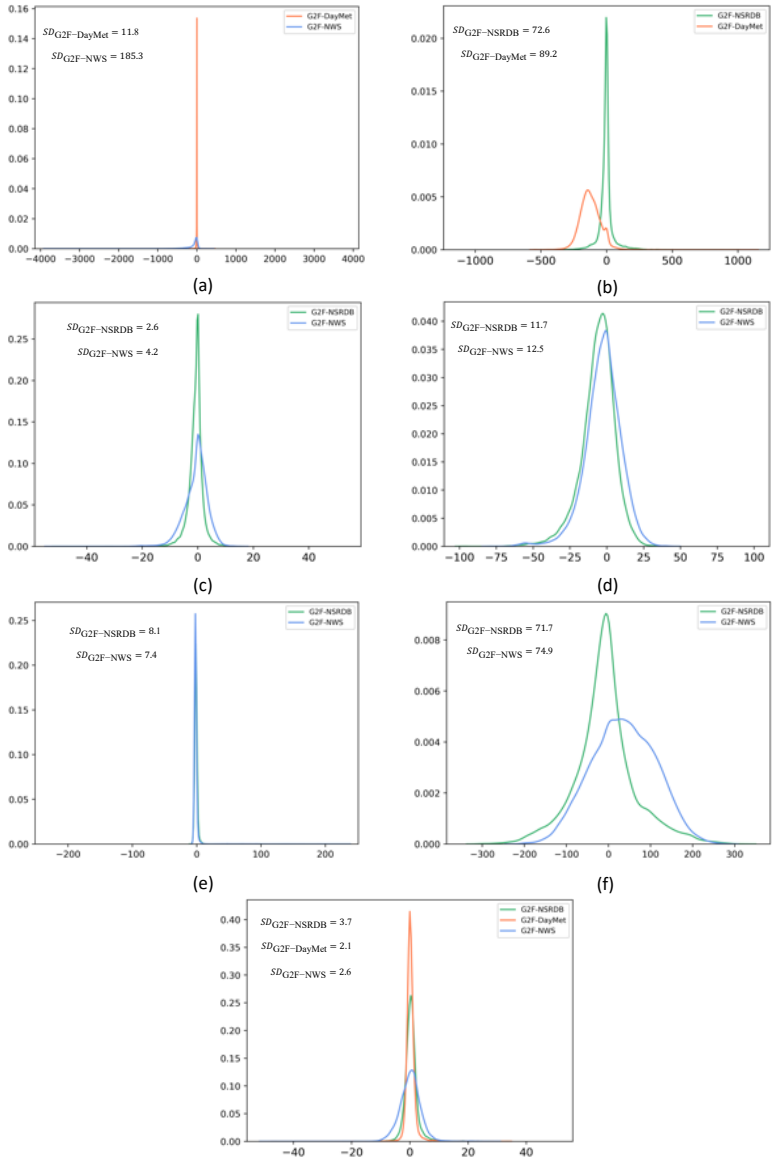
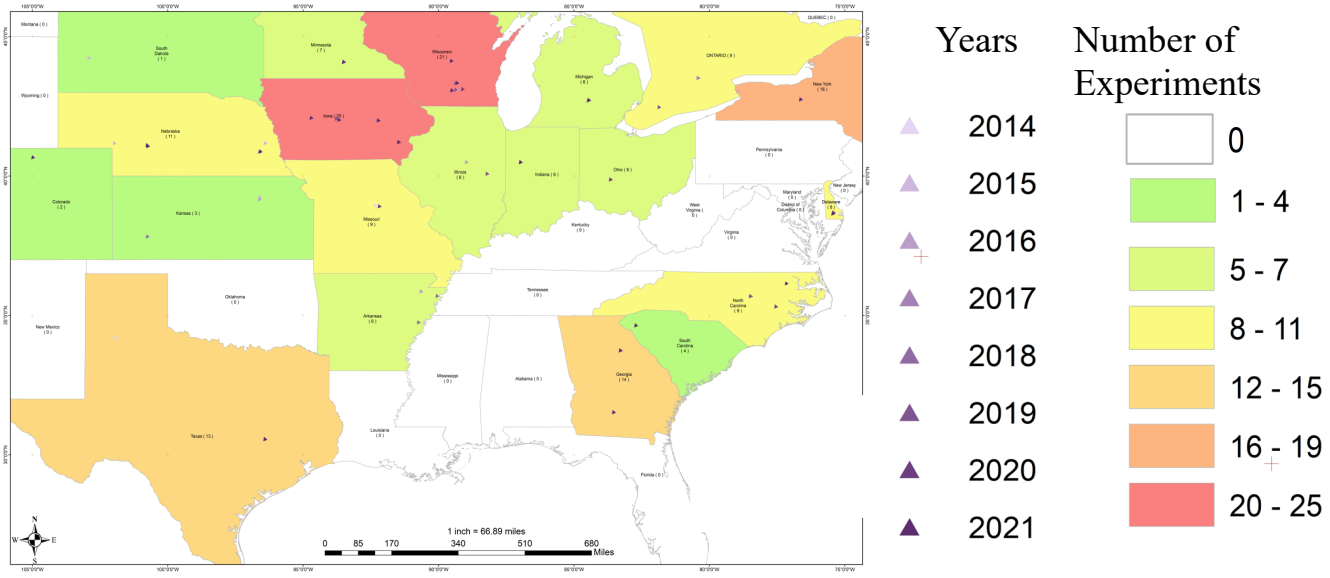


Figure 7. The number of observations of G2F-Gen. (genomic data), G2F-Phe. (phenotypic data), and G2F-Env. (environmental data) in the original database, quality-controlled database, and the consistency-controlled database. The QC and CC refer to quality and consistency control algorithms.



135

Figure 8. The probability distribution function (PDF) of the error values for (a) rainfall (Err-R), (b) solar radiation (Err-S), (c) dew point (Err-D), (d) relative humidity (Err-H), (e) wind speed (Err-W), (f) wind direction (Err-I), and (g) temperature (Err-T). Note that each of the external environmental data sources may not contain all the G2F hydroclimatic variables. The error term has been calculated for the common variables between G2F and each of the
140 **data sources. The SDG2F-NSRDB denotes the standard deviation of the errors between G2F and NSRDB, the SDG2F-DayMet denotes the standard deviation of the errors between G2F and DayMet, and SDG2F-NWS denotes the standard deviation of the errors between G2F and NWS for a given climatic variable.**



Supp. Figure 1. Locations, years, and number of sites per state used in by the G2F initiative and represented in the CLIM4OMICS.

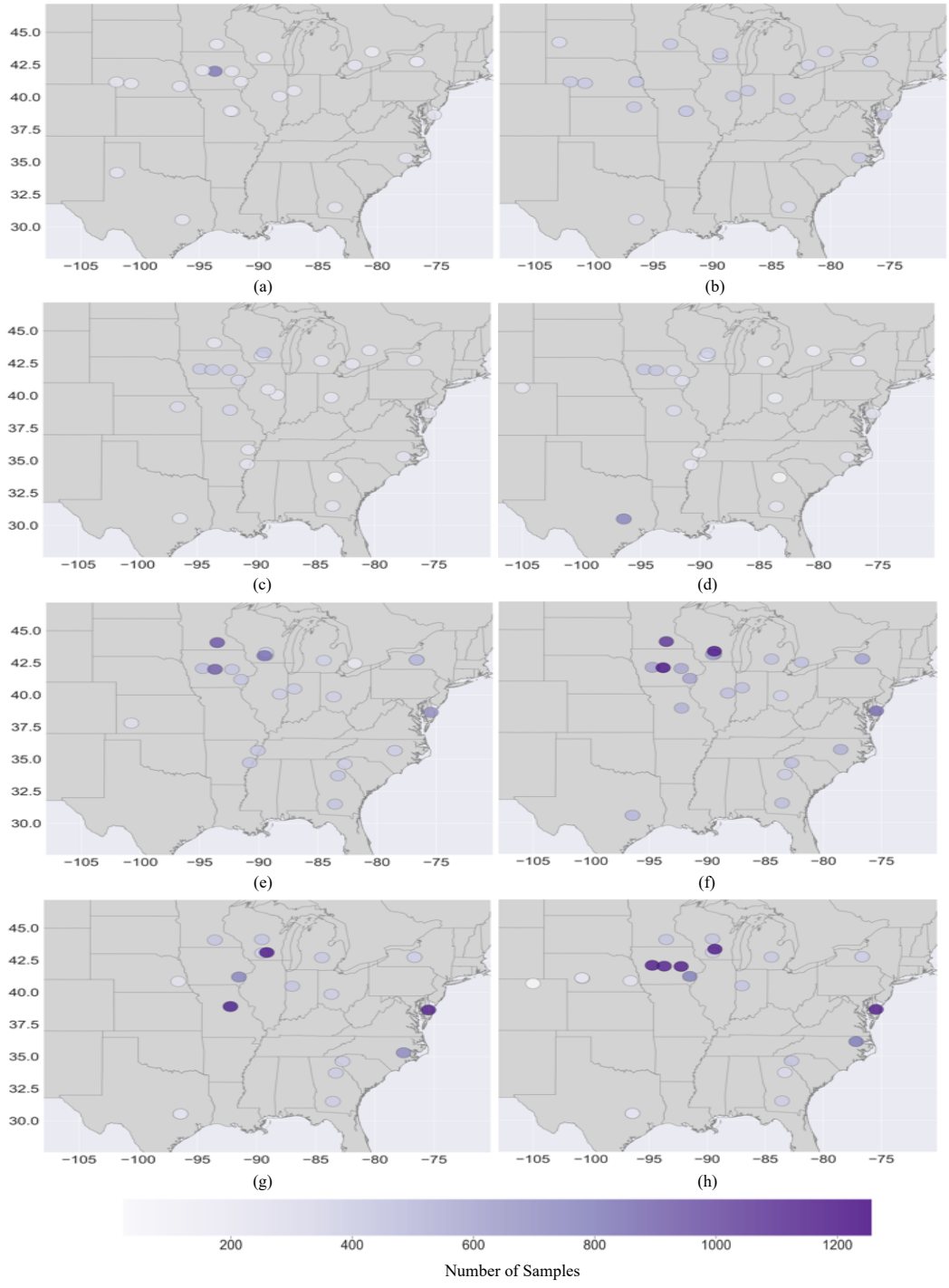
175

Figure 1. A conceptual framework of quality and consistency control algorithms for the multidimensional Genomes to Fields (G2F) OMICs and hydroclimatic database. "G2F-G" denotes G2F genomic data, "G2F-P" denotes G2F phenotypic data, "G2F-M" denotes G2F metadata, and "G2F-E" denotes G2F environmental data. The map indicates the locations and number of sites per state used in by the G2F initiative and represented in the CLIM4OMICS (the map is expanded as Supplementary Figure 1).

180

185

190



Supplementary Figure 2. The spatial distribution of phenotypic records in G2F-P database in (a) 2014 to (h) 2021.