What Does TERRA-REF's High Resolution, Multi Sensor Plant Sensing Public Domain Data Offer the Computer Vision Community?

David LeBauer
University of Arizona
dlebauer@arizona.edu

Max Burnette
University of Illinois
burnette@illinois.edu

Noah Fahlgren
Donald Danforth Plant Science Center
nfahlgren@danforthcenter.org

Rob Kooper
University of Illinois
kooper@illinois.edu

Kenton McHenry
University of Illinois
mchenry@illinois.edu

Abby Stylianou
St. Louis University
abby.stylianou@slu.edu

Abstract

A core objective of the TERRA-REF project was to generate an open-access reference dataset for the evaluation of sensing technologies to study plants under field conditions. The TERRA-REF program deployed a suite of high-resolution, cutting edge technology sensors on a gantry system with the aim of scanning 1 hectare ($10^4 m$) at around 1 mm^2 spatial resolution multiple times per week. The system contains co-located sensors including a stereo-pair RGB camera, a thermal imager, a laser scanner to capture 3D structure, and two hyperspectral cameras covering wavelengths of 300-2500nm. This sensor data is provided alongside over sixty types of traditional plant phenotype measurements that can be used to train new machine learning models. Associated weather and environmental measurements, information about agronomic management and experimental design, and the genomic sequences of hundreds of plant varieties have been collected and are available alongside the sensor and plant phenotype data.

Over the course of four years and ten growing seasons, the TERRA-REF system generated over 1 PB of sensor data and almost 45 million files. The subset that has been released to the public domain accounts for two seasons and about half of the total data volume. This provides an unprecedented opportunity for investigations far beyond the core biological scope of the project.

The focus of this paper is to provide the Computer Vision and Machine Learning communities an overview of the available data and some potential applications of this one of a kind data.

1. Introduction

In 2015, the Advanced Research Projects Agency for Energy (ARPA-E) funded the TERRA-REF Phenotyping Platform (Figure 1). The scientific aim was to transform plant breeding by providing a reference dataset generated by deploying a suite of co-located high-resolution sensors under field conditions. The goal of these sensors was to use proximate sensing from approximately 2m above the plant canopy to quantify plant characteristics.

The study has evaluated diverse populations of sorghum, wheat, and lettuce over the course of four years and ten cropping cycles. Future releases of additional data will be informed by user interests.

Figure 1. TERRA-REF field scanner at the University of Arizona's Maricopa Agricultural Center.

The TERRA-REF reference dataset can be used to characterize phenotype-to-genotype associations, on a genomic scale, that will enable knowledge-driven breeding and the development of higher-yielding cultivars of sorghum and wheat. The data is also being used to develop new algorithms for machine learning, image analysis, genomics, and optical sensor engineering. Beyond applications in plant

breeding, the resulting dataset provides opportunities for the study and integration of diverse remote sensing modalities.

1.1. Types of Data

The TERRA-REF field scanner platform utilizes a sensor suite of co-located instruments (Figure 2 and Table 1). The TERRA-REF reference dataset includes several data types (Figures 3 and 4, Table 2) including raw and processed outputs from sensors, environmental sensor measurements, manually measured and computationally derived phenotypes, and raw and processed genomics datasets [16]. Extensive contextual measurements and metadata include sensor information and extensive documentation for each of the sensors, the field scanner, calibration targets, and the results of sensor validation tests [16].

In addition to raw sensor data, the first release of TERRA-REF data includes derived sensor data products in enhanced formats including calibrated and georeferenced images and point clouds (Table 2). Many of the data products are provided in formats that follow Open Geospatial Consortium (OGC) standards and work with GIS software.
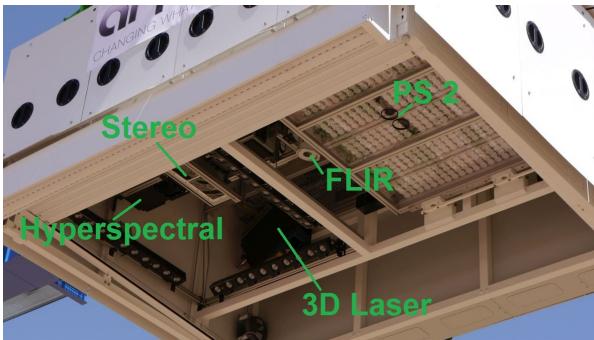

Figure 2. TERRA-REF field scanner sensor suite.

1.2. Sensors

Sensors available on the TERRA-REF field scanner include snapshot and line-scan imaging, multi-spectral radiometric, and environmental sensors. Table 1 and Figure 2) provide a high level overview of the sensors deployed on this system. Full documentation and metadata for each sensor as well as the configuration and geometry of the sensor box are provided as metadata alongside the TERRA-REF data release.

2. Computer Vision and Machine Learning Problems

There are a variety of questions that the TERRA-REF dataset could be used to answer that are of high importance to the agricultural and plant science communities, while also posing extremely interesting and challenging computer vision and machine learning problems. In this section, we consider example research areas or topics within computer vision and discuss the relevant agricultural and plant science questions those research communities could help address using the TERRA-REF data.

Measurement, Prediction and Causal Inference. The TERRA-REF sensor data can be used to drive development of vision-based algorithms for fundamental problems in plant phenotyping, such as making measurements of plant height, leaf length, flower counting, or estimating environmental stress. Additional challenges include attempting to predict *end of season* phenotypes, such as end of season yield, from early season sensor data – an accurate predictor of end of season yield from early season visual data, for example, could help growers and breeders invest resources only in the most promising of candidate crops. There are additional opportunities to investigate the causal relationship between genotypes or environmental conditions and their expressed phenotypes, as the TERRA-REF dataset includes both comprehensive genetic information, as well as high temporal resolution environmental information. The TERRA-REF data contain over sixty hand measurements that could be used to train models from one or more sensors. In addition, there are opportunities to train models that predict plot-level phenotypes measured by an expensive sensor with a less expensive sensor. Further, many events including insect damage, heat stress, and plant lodging (falling) could be labeled in new images.

Fine Grained Visual Categorization. The TERRA-REF data is a rich source of visual sensor data collected from crop species that are visually similar. Differentiating between data with low inter-class variance is an interesting categorization challenge, requiring visual models that learn the fine-grained differences between varieties of the same crop.

Transfer Learning. There are a variety of interesting transfer learning challenges of utmost importance to the agricultural and plant science communities, including discovering approaches that generalize across sensors, across crops, or across environmental conditions. The TERRA-REF data additionally presents an opportunity to help solve the greenhouse-to-field gap, where models that perform well in greenhouse conditions tend to not generalize to field conditions; because the TERRA-REF data includes both greenhouse and field data for the exact same varieties, researchers in transfer learning could help build models that bridge this gap.

Multi-sensor Integration. The TERRA-REF data includes data captured from a variety of visual sensors (described in Section 1.2). These sensors have similar, but not

Table 1. Summary of TERRA-REF sensor instruments.

Sensor Name	Model	Technical Specifications
Imaging Sensors		
Stereo RGB Camera	Allied Vision Prosilica GT3300C	
Laser Scanner	Custom Fraunhofer 3D	Spatial Resolution: 0.3 to 0.9 mm
Thermal Infrared	FLIR A615	Thermal Sensitivity: $\leq$50mK @ 30 °C
PS II Camera	LemnaTec PS II Fluorescence Prototype	Illumination 635nm x 4000 μmol/m^2/s, Camera 50 fps
Multi-spectral Radiometers		
Dedicated NDVI Multispectral Radiometer	Skye Instruments SKR 1860D/A	650 nm, 800 nm $\pm$ 5 nm; 1 down, 1 up
Dedicated PRI Multispectral Radiometer	Skye Instruments SKR 1860ND/A	531nm +/- 3nm; PRI = Photochemical Reflectance Index
Active Reflectance	Holland Scientific Crop Circle ACS-430	670 nm, 730 nm, 780 nm
Hyper-spectral Cameras		
VNIR Hyperspectral Imager	Headwall Inspector VNIR	380-1000 nm @ 2/3 nm resolution
SWIR Hyperspectral Imager	Headwall Inspector SWIR	900-2500 nm @ 12 nm resolution
Environmental Sensors		
Climate Sensors	Thies Clima 4.9200.00.000	
VNIR Spectroradiometer	Ocean Optics STS-Vis	Range: 337-824 nm @ 1/2 nm
VNIR+SWIR Spectroradiometer	Spectral Evolution PSR+3500	Range 800-2500nm @3-8 nm; Installed 2018
PAR Sensor	Quantum SQ–300	Spectral Range 410 to 655 nm

Table 2. Summary of the sensor data products included in the first release of TERRA-REF data.

Data Product	Sensor	Algorithm	File Format	Plot Clip	Full Field
Environment	Thies Clima	envlog2netcdf	netcdf	NA	NA
Thermal Image	FLIR	ir_geotiff	geotiff	+	
Point Cloud	Fraunhofer Laser 3D	laser3d_las	las	+	
Point Cloud	Fraunhofer Laser 3D	scanner3DTop	ply		
Images Time-Series	PSII Camera	ps2png	png		
Color Images	RGB Stereo	bin2tiff	geotiff	+	+
Plant Mask	RGB Stereo	rgb_mask	geotiff		x

identical, viewpoints from within the gantry box, may not have captured data at the exact same time, and may have captured different perspectives of the same part of the field on different days. This presents interesting challenges in terms of how to incorporate information across the various sensors, and how to work with time-series data that is not necessarily well-aligned or continuously captured.

Explainable Models. All too often in machine learning research, datasets and models are built solely to drive the development of machine learning algorithms. When building models to answer questions like "should I cut this plant down because it won't produce sufficient yield?" or "is this plant under environmental stress?," it is important not just to have maximally accurate models but to also understand *why* the models make the determinations that they make. This makes the TERRA-REF data, and the biologically relevant questions it supports, an excellent opportunity to drive development of new approaches for explainable machine learning, conveying the decisions made by algorithms to non-machine learning experts.

Information Content. The TERRA-REF field scanner and sensors represent a substantial investment, and it is still not clear which sensors, sensor configurations, and spatial and temporal resolutions are useful to answer a particular question. Presently, much less expensive sensors and sensing platforms are available [11, 1]. What do we gain from the 1mm spatial resolution on this platform relative to unoccupied aerial systems (UAS) that are quickly approaching 1cm spatial resolution? Or, which subset of hyperspectral wavelengths provide the most useful information? Can we predict the useful parts of a hyperspectral image from RGB images? Or get most of the information from a multispectral camera with a half-dozen bands? At the outset, the team recognized that this configuration would be oversampling the plant subjects, but it wasn't clear what the appropriate resolutions or most useful sensors would be.

Overall Challenges. Within all of these topic areas in computer vision and machine learning, the challenges that must be addressed require addressing interesting questions such as determining the most appropriate sensors and data processing choices for specific questions, addressing difficult domain transfer issues, considering how to integrate noisy side channels of information, such as genetic information that may conflict or conflate with each other, or dealing with nuisance parameters like environmental or weather variations that simultaneously influence plant subjects and the sensor data content.
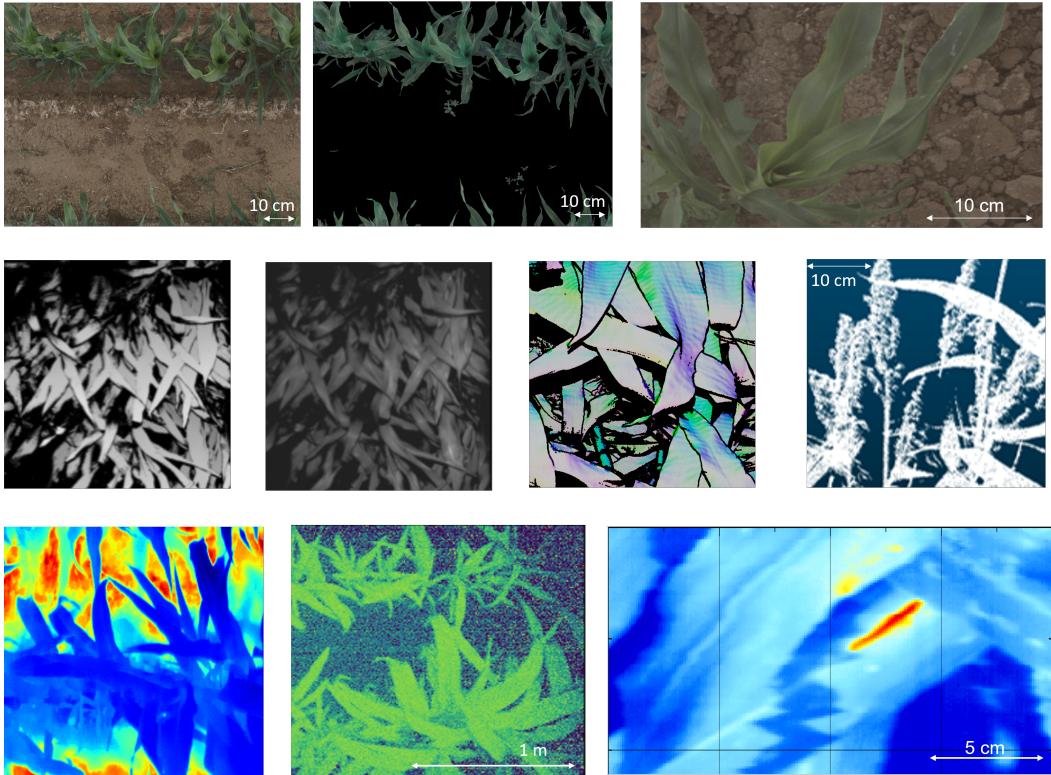
Figure 3. **Example data from the TERRA-REF gantry system.** (top left) RGB data (center top) RGB data with soil masked (top right) close up of RGB data. 3D-scanner data (center row, right to left) depth, reflectance, surface normals, and point cloud data produced by the 3D scanner. (bottom row, right to left) FLIR thermal image with transpiring leaves shown as cooler than soil; F_v/F_m derived from active fluorescence PSII sensor providing a measure of photosynthetic efficiency; the reflectance of light at 543nm wavelength measured by the VNIR hyperspectral camera. Because these are two-dimensional representations of three-dimensional systems, all scale bars are approximate.

3. Algorithm Development.

The process of converting raw sensor outputs into usable data products required geometric, radiometric, and geospatial calibration. In this regard, each sensor presented its own challenges. Combining these steps into an automated computing pipeline also represented a substantial effort that is described by Burnette *et al*. [3].

Radiometric calibration was particularly challenging, owing that many images contain both sunlit and shaded areas. In the case of hyperspectral images, the white sensor box and scans spread out over multiple days confounded an already challenging problem. Radiometric calibration of images taken by the two hyperspectral cameras exemplifies these challenges, and a robust solution is described by Sagan *et al*. [21] and implemented in [19]. Even processing images from an RGB camera was challenging due to fixed settings resulting in high variability in quality and exposure, requiring the novel approach described by Li *et al*. [18]. Herritt *et al*. [14, 13] demonstrate and provide software used in analysis of a sequence of images that capture plant fluorescence response to a pulse of light.

Most of the algorithms used to generate data products have not been published as papers but are made available on GitHub (https://github.com/terraref); code used to release the data publication in 2020 is available on Zenodo [25, 15, 10, 6, 4, 19, 8, 7, 5, 9, 17].

Pipeline development continues to support ongoing use of the field scanner as well as more general applications in plant sensing pipelines. Recent advances have improved pipeline scalability and modularity by adopting workflow tools and making use of heterogeneous computing environments. The TERRA-REF computing pipeline has been adapted and extended for continuing use with the Field Scanner with the new name "PhytoOracle" and is available at https://github.com/LyonsLab/PhytoOracle. Related work generalizing the pipeline for other phenomics applications has been released under the name "AgPipeline" https://github.com/agpipeline with applications to aerial imaging described by Schnaufer *et al*. [22]. All of these software are made available with permissive open source licenses on GitHub to enable access and community development.
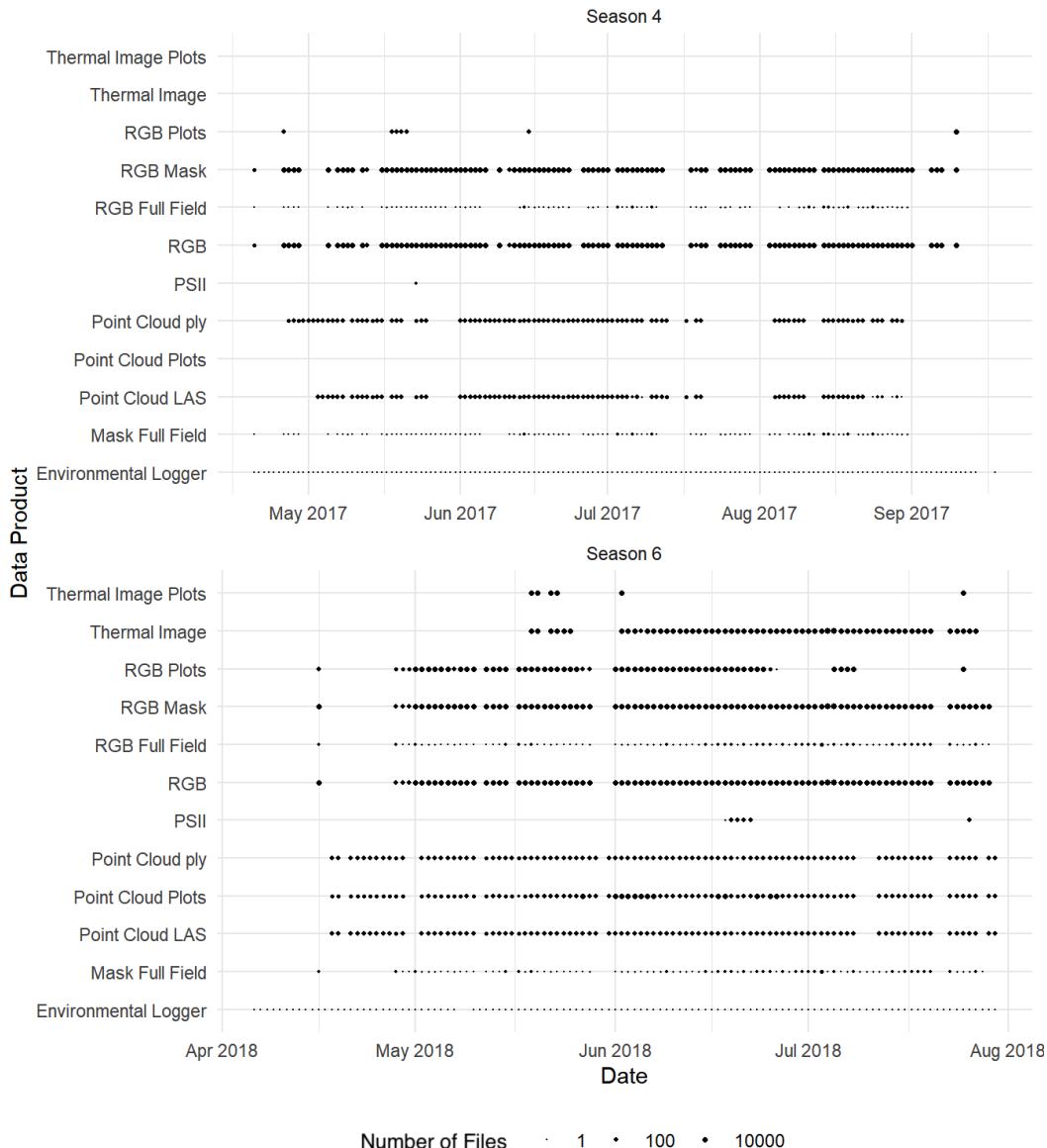
Figure 4. Summary of public sensor datasets from Seasons 4 and 6. Each dot represents the dates for which a particular data product is available, and the size of the dot indicates the number of files available.

4. Uses to date.

TERRA-REF is being used in a variety of ways. For example, hyperspectral images have been used to measure soil moisture [2], but the potential to predict leaf chemical composition and biophysical traits related to photosynthesis and water use are particularly promising based on prior work [23, 24].

Plant Science and Computer Vision Research. A few projects are developing curated datasets for specific machine learning challenges related to classification, object recognition, and prediction.

We currently know of at least three datasets curated for CVPPA 2021. The Sorghum-100 dataset was created to support development of algorithms that can classify sorghum varieties from RGB images from Ren *et al*. [20]. Another set of RGB images curated for the Sorghum Biomass Prediction Challenge on Kaggle was developed with the goal of developing methods to predict end of season biomass from images taken of different sorghum genotypes over the course of the growing season. Finally, RGB images from the TERRA-REF field scanner in Maricopa accounted for 250 of the 6000 1024x1024 pixel images in

the Global Wheat Head Dataset 2021 [12]. The goal of the Global Wheat Challenge 2021 on AIcrowd is to develop an algorithm that can identify wheat heads from a collection of images from around the world that represent diverse fields conditions, sensors, settings, varieties, and growth stages.

Most of the research applications to date have focused on analysis of plot-level phenotypes and genomic data rather than the full resolution sensor data.

5. Data Access

Public Domain Data. A curated subset of the TERRA-REF data was released to the public domain in 2020 (Figure 4) [16]. These data are intended to be re-used and are accessible as a combination of files and databases linked by spatial, temporal, and genomic information. In addition to providing open access data, the entire computational pipeline is open source, and we can assist academic users with access to high-performance computing environments.

The total size of raw (Level 0) data generated by these sensors is 60 TB. Combined, the Level 1 and Level 2 sensor data products are 490 TB. This size could be substantially reduced through compression and removal of duplicate data. For example, the same images at the same resolution appear in the georeferenced Level 1 files, the full field mosaics, and the plot-level clip.

Other Data Available. The complete TERRA-REF dataset is not publicly available because of the effort and cost of processing, reviewing, curating, describing, and hosting the data. Instead, we focused on an initial public release and plan to make new datasets available based on need. Access to unpublished data can be requested from the authors, and as data are curated they will be added to subsequent versions of the public domain release (https://terraref.org/data/access-data).

In addition to hosting an archival copy of data on Dryad [16], the documentation includes instructions for browsing and accessing these data through a variety of online portals. These portals provide access to web user interfaces as well as databases, APIs, and R and Python clients. In some cases it will be easier to access data through these portals using web interfaces and software libraries.

The public domain data is archived on Dryad, with the exception of the large sensor data files. The Dryad archive provides a catalog of these files that can be accessed via Globus or directly on the host computer at the National Center for Supercomputing Applications.

6. Acknowledgements

References

[1] Jonathan A Atkinson, Robert J Jackson, Alison R Bentley, Eric Ober, and Darren M Wells. *Field Phenotyping for the Future*, pages 1–18. John Wiley & Sons, Ltd, Chichester, UK, Nov. 2018. 3

[2] Ebrahim Babaeian, Paheding Sidike, Maria S Newcomb, Maitiniyazi Maimaitijiang, Scott A White, Jeffrey Demieville, Richard W Ward, Morteza Sadeghi, David S LeBauer, Scott B Jones, et al. A new optical remote sensing technique for high-resolution mapping of soil moisture. *Frontiers in Big Data*, 2:37, 2019. 5

[3] Maxwell Burnette, Rob Kooper, J D Maloney, Gareth S Rohde, Jeffrey A Terstriep, Craig Willis, Noah Fahlgren, Todd Mockler, Maria Newcomb, Vasit Sagan, Pedro Andrade-Sanchez, Nadia Shakoor, Paheding Sidike, Rick Ward, and David LeBauer. TERRA-REF data processing infrastructure. In *Proceedings of the Practice and Experience on Advanced Research Computing*, page 27. ACM, July 2018. 4

[4] Max Burnette, David LeBauer, Solmaz Hajmohammadi, ZongyangLi, Craig Willis, Wei Qin, Sidke Paheding, and JD Maloney. terraref/extractors-multispectral: Season 6 Data Publication (2019), Sept. 2019. 4

[5] Max Burnette, David LeBauer, Wei Qin, and Yan Liu. terraref/extractors-metadata: Season 6 Data Publication (2019), Sept. 2019. 4

[6] Max Burnette, David LeBauer, ZongyangLi, Wei Qin, Solmaz Hajmohammadi, Craig Willis, Sidke Paheding, and Nick Heyek. terraref/extractors-stereo-rgb: Season 6 Data Publication (2019), Sept. 2019. 4

[7] Max Burnette, Zongyang Li, Solmaz Hajmohammadi, David LeBauer, Nick Heyek, and Craig Willis. terraref/extractors-3dscanner: Season 6 Data Publication (2019), Sept. 2019. 4

[8] Max Burnette, Jerome Mao, David LeBauer, Charlie Zender, and Harsh Agrawal. terraref/extractors-environmental: Season 6 Data Publication (2019), Sept. 2019. 4

[9] Max Burnette, Craig Willis, Chris Schnaufer, David LeBauer, Nick Heyek, Wei Qin, Solmaz Hajmohammadi, and Kristina Riemer. terraref/terrautils: Season 6 Data Publication (2019), Sept. 2019. 4

[10] Max Burnette, Charlie Zender, JeromeMao, David LeBauer, Rachel Shekar, Noah Fahlgren, Craig Willis, Henry Butowsky, Xingchen Hong, ZongyangLi, Fengling Wang, TinoDornbusch, JD Maloney, Wei Qin, Stuart Marshall, Abby Stylianou, and Ting Li. terraref/computing-pipeline: Season 4 & 6 Data Publication (2019), Feb. 2020. 4

[11] Anna L Casto, Haley Schuhl, Jose C Tovar, Qi Wang, Rebecca S Bart, Noah Fahlgren, and Malia A Gehan. Picturing the future of food. *Plant phenome j.*, 4(1), Jan. 2021. 3

[12] Etienne David, Mario Serouart, Daniel Smith, Simon Madec, Kaaviya Velumani, Shouyang Liu, Xu Wang, Francisco Pinto Espinosa, Shahameh Shafiee, Izzat S. A. Tahir,

Hisashi Tsujimoto, Shuhei Nasuda, Bangyou Zheng, Norbert Kichgessner, Helge Aasen, Andreas Hund, Pouria Sadhegi-Tehran, Koichi Nagasawa, Goro Ishikawa, Sébastien Dandrifosse, Alexis Carlier, Benoit Mercatoris, Ken Kuroki, Haozhou Wang, Masanori Ishii, Minhajul A. Badhon, Curtis Pozniak, David Shaner LeBauer, Morten Lilimo, Jesse Poland, Scott Chapman, Benoit de Solan, Frédéric Baret, Ian Stavness, and Wei Guo. Global wheat head dataset 2021: more diversity to improve the benchmarking of wheat head localization methods, 2021. 6

[13] Matthew T Herritt, Jacob C Long, Mike D Roybal, David C Moller Jr, Todd C Mockler, Duke Pauli, and Alison L Thompson. Flip: Fluorescence imaging pipeline for field-based chlorophyll fluorescence images. *SoftwareX*, 14:100685, 2021. 4

[14] Matthew T Herritt, Duke Pauli, Todd C Mockler, and Alison L Thompson. Chlorophyll fluorescence imaging captures photochemical efficiency of grain sorghum (sorghum bicolor) in a field setting. *Plant methods*, 16(1):1–13, 2020. 4

[15] David LeBauer, Nick Heyek, Rachel Shekar, Katrin Leinweber, JD Maloney, and Tino Dornbusch. terraref/reference-data: Season 4 & 6 Data Publication (2019), Feb. 2020. 4

[16] David LeBauer, Burnette Maxwell, Jeffrey Demieville, Noah Fahlgren, Andrew French, Roman Garnett, Zhenbin Hu, Kimberly Huynh, Rob Kooper, Zongyang Li, Maitiniyazi Maimaitijiang, Jerome Mao, Todd Mockler, Geoffrey Morris, Maria Newcomb, Michael Ottman, Philip Ozersky, Sidike Paheding, Duke Pauli, Robert Pless, Wei Qin, Kristina Riemer, Gareth Rohde, William Rooney, Vasit Sagan, Nadia Shakoor, Abby Stylianou, Kelly Thorp, Richard Ward, Jeffrey White, Craig Willis, and Charles Zender. Data from: TERRA-REF, an open reference data set from high resolution genomics, phenomics, and imaging sensors, Aug. 2020. 2, 6

[17] David LeBauer, Craig Willis, Rachel Shekar, Max Burnette, Ting Li, Scott Rohde, Yan Liu, JD Maloney, Noah Fahlgren, Charlie Zender, Rob Kooper, Jerome Mao, Harsh Agrawal, Xingchen Hong, Shannon Bradley, Samy Pessé, Katrin Leinweber, Justin Manzo, Jeff Terstriep, and Abby Stylianou. terraref/documentation: Season 6 Data Publication (2019), Feb. 2020. 4

[18] Zongyang Li, Abby Stylianou, and Robert Pless. Learning to correct for bad camera settings in large scale plant monitoring. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, June 2019. 4

[19] Jerome Mao, Max Burnette, Henry Butowsky, Charlie Zender, David LeBauer, and Sidke Paheding. terraref/extractors-hyperspectral: Season 6 Data Publication (2019), Sept. 2019. 4

[20] Chao Ren, Justin Dulay, Gregory Rolwes, Duke Pauli, Nadia Shakoor, and Abby Stylianou. Multi-resolution outlier pooling for sorghum classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pages 2931–2939, June 2021. 5

[21] Vasit Sagan, Maitiniyazi Maimaitijiang, Sidke Paheding, Sourav Bhadra, Nichole Gosselin, Max Burnette, Jeffrey Demieville, Sean Hartling, David LeBauer, Maria Newcomb, Duke Pauli, Kyle T. Peterson, Nadia Shakoor, Abby Stylianous, Charlie Zender, and Todd C. Mockler. Data-driven artificial intelligence for calibration of hyperspectral big data. *Transactions on Geoscience and Remote Sensing*, *2021*. 4

[22] Christophe Schnaufer, Julian L Pistorius, and David S LeBauer. An open, scalable, and flexible framework for automated aerial measurement of field experiments. In *Proceedings of SPIE*, 2020. 4

[23] Shawn P Serbin, Dylan N Dillaway, Eric L Kruger, and Philip A Townsend. Leaf optical properties reflect variation in photosynthetic metabolism and its sensitivity to temperature. *J. Exp. Bot.*, 63(1):489–502, Jan. 2012. 5

[24] Shawn P Serbin, Aditya Singh, Ankur R Desai, Sean G Dubois, Andrew D Jablonski, Clayton C Kingdon, Eric L Kruger, and Philip A Townsend. Remotely estimating photosynthetic capacity, and its response to temperature, in vegetation canopies using imaging spectroscopy. *Remote Sens. Environ.*, 167:78–87, Sept. 2015. 5

[25] Craig Willis, David LeBauer, Max Burnette, and Rachel Shekar. terraref/sensor-metadata: Season 4 & 6 Data Publication (2019), Feb. 2020. 4