Chapter 6

Data Interoperability Tools for Regional Integrated Assessments

Cheryl H. Porter[1], Chris Villalobos[1], Dean Holzworth[2], Roger Nelson[3], Jeffrey W. White[4], Ioannis N. Athanasiadis[5], Meng Zhang[1], Sander Janssen[6], Rob Knapen[6], James W. Jones[1], Kenneth J. Boote[1], John Hargreaves[2], and John M. Antle[7]

[1]*University of Florida, Gainesville, FL, USA*

[2]*Commonwealth Scientific and Industrial Research Organisation, Toowoomba, Australia*

[3]*Washington State University, Pullman, WA, USA*

[4]*United States Department of Agriculture, Agricultural Research Service, Ames, IA, USA*

[5]*Democritus University of Thrace, Xanthi, Greece*

[6]*Alterra, Wageningen UR, the Netherlands*

[7]*Oregon State University, Corvallis, OR, USA*

Introduction

The crop modeling simulations performed as part of the regional integrated assessments (RIA) for AgMIP (Agricultural Model Intercomparison and Improvement Project; Rosenzweig *et al.*, 2013a) involve the use of multiple models, which allows researchers to characterize uncertainty and to facilitate model intercomparisons and improvement. The emphasis on ensemble modeling approaches requires that data used in multi-model simulations be harmonized across models. Common data descriptions for field conditions, management operations, soils characteristics, and weather data must be provided equivalently to each model so that differences in simulated outputs are due to model differences rather than input data differences, as much as possible. This chapter describes the AgMIP approach to achieving data interoperability across multiple crop models, which consists of establishing an efficient standardized data-exchange mechanism with specifications defined in accordance with international standards; implementing a flexibly structured data schema to store experimental datasets; and providing consistent procedures for supplementing model-required inputs (Porter *et al.*, 2014).

The AgMIP RIA crop modeling process begins with collection of data required to parameterize the crop models. Field-management conditions are typically obtained from relatively low-quality data sources such as farm surveys. These data are supplemented with knowledge from regional agronomists regarding typical practices such as cultivars grown and crop-residue management. Soil-profile data are obtained from various sources including the Harmonized World Soils Database (HWSD; FAO/IIASA/ISRIC/ISSCAS/JRC, 2012) and the WISE soil database (Batjes, 2009), complemented with information gained from farm surveys and from knowledge of cropping history. Daily weather data used to drive the models are generated by the AgMIP climate research teams, as detailed in Part 1, Chapter 3 in this volume. The RIA crop modeler transforms these data into fully and correctly parameterized model inputs which describe the management and environmental conditions for the observed farming system using methods described later in this chapter. Model simulations representing these historic or observed conditions are then used to compute model biases.

The historic simulations are used as a basis to extrapolate predictions of crop production for 30 years using: (1) current climate and current management, (2) future climate and current management, and (3) future climate and adapted management. Simulations are performed for multiple climate models, time-periods, emissions scenarios, socio-economic scenarios, and management strategies for climate change adaptation.

Consider a typical case in which an AgMIP regional research team (RRT) is analyzing four regions, each with 150 survey sites over a 30-year time-frame. Suppose historical climate plus four climate change scenarios, generated with five climate models are simulated by using two crop models. This would result in over 750,000 simulations. One must also consider that there will inevitably be procedural mistakes and data corrections requiring multiple re-runs of simulations. Adaptation scenarios and additional crop or climate models add further dimensions to the simulation combinatorics. It is clear that automation is required to facilitate the generation of simulation scenarios, translation of data for multiple crop models, aggregation of outputs, analysis, and visualization of results. This was the motivation for AgMIP to develop a set of standards and protocols for crop model data management and interoperability for RIAs.

Some of the challenges encountered with design of data interoperability protocols involve (1) the use of heterogeneous data from disparate sources, (2) uniformly supplying the required model assumptions where data are insufficient to parameterize the crop models, (3) supplying observed and supplemental data equivalently to multiple models each with different input formats, and (4) harmonizing model outputs for use in further analyses. Figure 1 illustrates these challenges to standardizing AgMIP crop modeling data and streamlining data processing for multiple

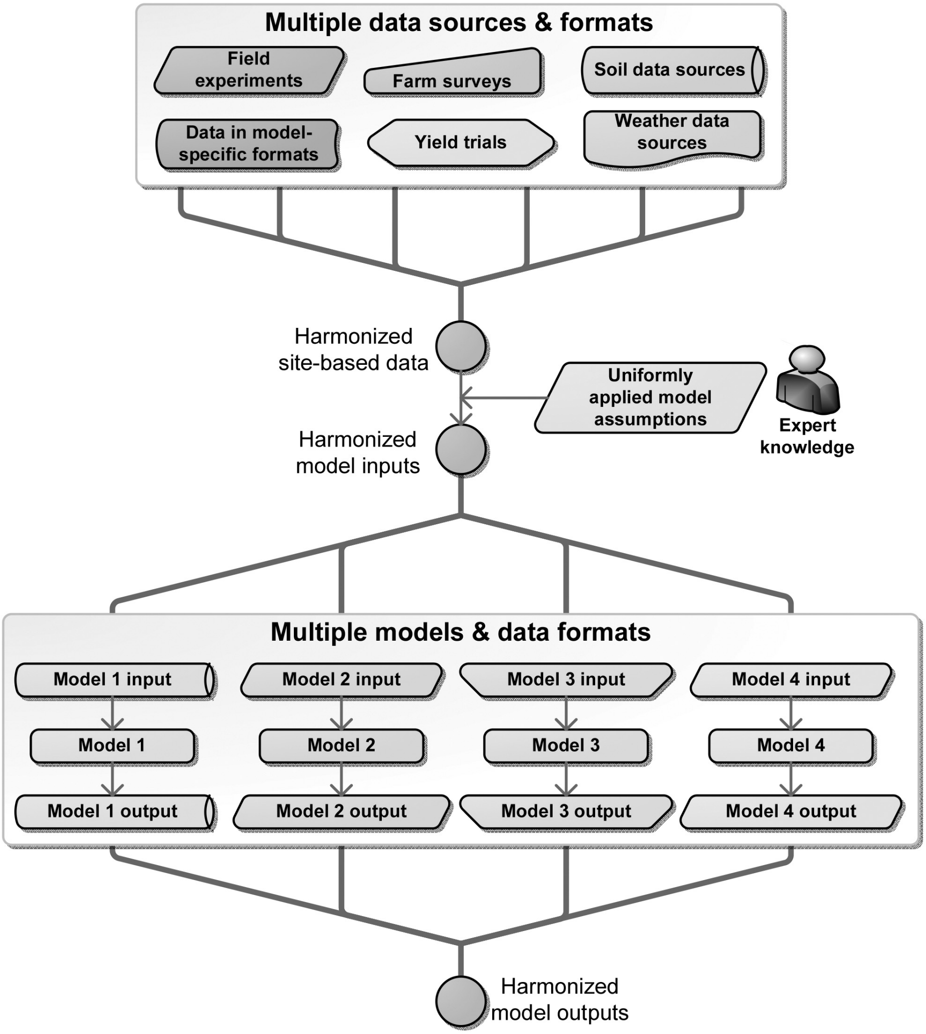

Fig. 1. Data-flow diagram for AgMIP crop modeling activities. Source data, model inputs, and model outputs have diverse and inconsistent formats, hampering ensemble modeling approaches (Porter *et al.*, 2014). Supplemental data and assumptions must be provided equivalently to multiple models. Reprinted from Porter *et al.* (2014), with permission from Elsevier.

models and the mechanisms used to overcome them. Details of the AgMIP data interoperability solution are provided in the next section.

The first challenge deals with inconsistency in the **source data** used in crop modeling. Datasets used to calibrate and validate crop models are obtained from diverse sources and the quantity and quality of the data are influenced by the original purpose of the research, the equipment used for measurement of growth and environmental data, and the methods of analysis used. Field experiments typically collect detailed records of soil properties, weather conditions, and management practices such as tillage, fertilizer, and water-application dates, rates, methods, and materials. However, farm and household surveys may only collect general information about management practices with only seasonal totals and few details regarding management events. Field studies often record details about crop phenology, canopy, and biomass development and response to management practices, whereas such information is often lacking in surveys. These data sources also vary in storage and schema implementation formats, including databases, spreadsheets, tabular text files, structured XML (extensible markup language), or other specialized formats. Adoption of a common vocabulary and a standard data schema enable AgMIP researchers to overcome these source data incompatibility problems.

The second challenge in AgMIP data harmonization involves the issue of providing a consistent set of **assumptions** for information that is required by models but is not available in the datasets. Even highly detailed field experiments may lack sufficient observations to adequately provide minimum input requirements to the models. For ensemble modeling, it is necessary to harmonize not only the recorded information but also the assumptions used to supply missing model-specific input parameters. This challenge was overcome by development of a system for applying model assumptions to datasets based on the best available information for the region. This system, referred to as the DOME (Data Overlay for Multi-model Export), is discussed below.

The third challenge in data harmonization involves the inconsistent **input data** formats used by crop models. While the various models may implement diverse algorithms, the driving data are generally similar (Boote *et al.*, 2013). Models typically include daily weather records, soil physical and chemical properties, and information related to management practices such as planting, fertilization, irrigation, and harvest. However, the formats in which these data are provided to the models vary widely. This could be a matter of design choice, preference, availability, or even understanding of the underlying phenomena. As an example, soil-data specifications may differ in the soil classification used (sometimes incompatible with each other) and in how vertical differences in a given soil profile are specified (e.g., number and depth of soil layers). A library of model input translators was developed to overcome this challenge.

Likewise, although the core **outputs** from crop models are often similar, the formats are diverse and inconsistent and generally parallel the input-format structure. For example, models running within the Decision Support System for Agrotechnology Transfer (DSSAT; Jones *et al.*, 2003; Hoogenboom *et al.*, 2010) rely heavily on column formatted text for both input and output files, while the Agricultural Production Systems Simulator (APSIM; Holzworth *et al.*, 2013; Keating *et al.*, 2003) relies on XML files, with a semantic heterogeneity on the variable names and definitions. Model output translators were developed to harmonize the outputs from each participating model.

The key elements of the crop model data interoperability system for AgMIP are a common vocabulary for standardization of definitions and units; a means of harmonizing data collected from disparate sources and which differ considerably in content and quality; a means of providing datasets in the model-ready formats required by multiple models; and a means of harmonizing simulated model outputs. Harmonized datasets must be searchable with source data and model outputs linked in such a way that model results can be documented and replicated.

AgMIP Data Interoperability Solution

AgMIP crop experiment (ACE) harmonized data schema

The AgMIP crop experiment (ACE) harmonized format was designed to provide a means of storing the highly variable types of data associated with AgMIP crop modeling exercises in a flexible and efficient schema. The ACE data format relies on a common vocabulary that is independent of the structure and level of detail provided by the data sources. The standards developed by the International Benchmark Sites Network for Agrotechnology Transfer Project (IBSNAT, ICRISAT, 1984; Uehara and Tsuji, 1998) and subsequently revised by the International Consortium for Agricultural Systems Applications (ICASA) were designed to provide a comprehensive means of describing field experiments with the goal of facilitating exchange of agricultural information (Hunt *et al.*, 2001; White *et al.*, 2013).

The foundation of the ICASA standards is the Master Variables List (ICASA-MVL, see http://research.agmip.org/display/it/Data+Interoperability). This data dictionary is organized in a hierarchical structure with four major divisions: management practices, soil characteristics, weather, and measurements of crop and soil responses. The ICASA-MVL was selected as the foundation of the ACE data definitions because of its comprehensive and detailed descriptions of management practices and traits of soils and plants for crop experiments. The list is flexible and extensible, which allows AgMIP researchers to modify and improve the list as needed for AgMIP modeling activities. These changes included defining model variables that were not previously described, clarifying definitions that could have distinct interpretation in different models, and accommodating metadata specifically required for AgMIP purposes.

AgMIP site-based agricultural datasets have a wide range of quality and quantity of records collected depending on the data sources, and therefore these data do not easily conform to a rigid schema. A flexible, non-relational database architecture was selected to accommodate the diverse data structures and contents encountered.

Data are organized in ACE using type-agnostic JSON (JavaScript object notation, www.json.org) key-value structures. These key-value structures store all data in <attribute name, value> pairs. This allows the storage of information for both simple observations and complex structured objects in an efficient, consistent and open-ended format. The key (attribute name) in each pair corresponds to an ICASA-MVL variable; the value conforms to the variable definition, including units.

The core structure of ACE divides model data into three compartments that reflect the common drivers in most models, i.e., (a) experimental management, (b) soil characteristics, and (c) weather-related information. Each compartment is assigned an identification key, or hash code (NIST, 2012), generated from the contents. Weather and soil structures are associated with experimental site data via these identification keys. This allows imposition of one-to-many relations between experimental data and soils or weather data.

An important feature of ACE data is that there are no minimum data other than location (latitude and longitude) and crop species. All other data are optional and the data structure will contain key-value pairs for only the data collected in the survey or experiment. This provides a convenient and efficient data-storage structure.

Figure 2 presents two sample JSON fragments for an AgMIP dataset which includes four soil nodes (numbered 0–3), five experiment nodes, and four weather nodes. In Fig. 2a, the second field-experiment data node is expanded to show field geolocation, metadata, and the initial field conditions and management substructures. The "fl_lat" key corresponds to the ICASA variable defining field latitude as 37.75°S. One of the three management event substructures is expanded to show a fertilizer event with 120 kg/ha urea added (keys "feamn" and "fecd"), broadcast and incorporated to a depth of 5 cm (keys "feacd" and "fedep"). Figure 2b shows an expansion of the data in one of the soil-data structures, including one of seven soil layers. The key "sloc", for example, corresponds to a soil organic-carbon value of 0.31%. The soil for the experiment site, identified by "WP_ARG1234" (highlighted in Fig. 2a), is associated with the soil definition in the experiment data (shown in Fig. 2b) via the "soil_id" key field.

Data Overlay for Multi-model Export (DOME)

DOME provides a mechanism for users to provide model-specific assumptions and has several potential uses for model intercomparison. First, this feature allows researchers to provide supplemental information in order to utilize incomplete datasets. Missing parameters that are required by models can be supplied through one or more DOMEs which allow researchers to make assumptions uniformly based on the best agronomic knowledge of cultural practices in a region. These supplemental data are applied as an overlay, which allows the assumed information to remain

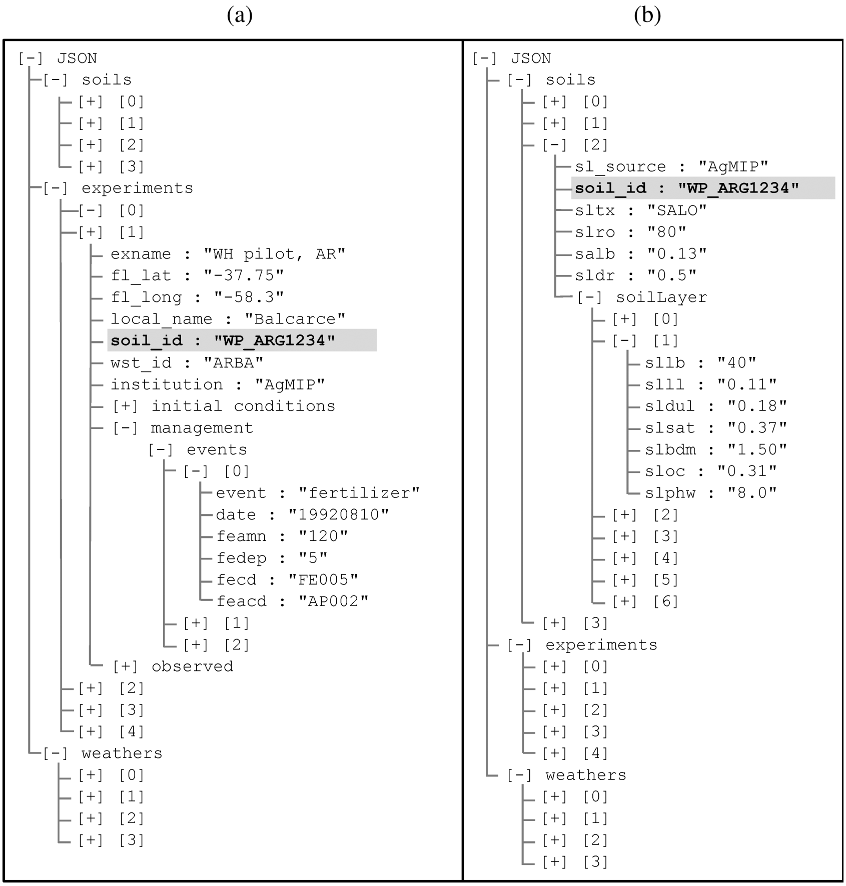

Fig. 2. Fragment of ACE JSON structure showing (a) expanded experiment data and (b) expanded soil data. The highlighted soil identifiers show how a specific soil profile is linked to an experiment.

separate from field-measured observations and ensures that all models are provided with a consistent set of assumptions. A DOME which is specifically used to provide a complete dataset for crop model simulations is referred to as a "*field overlay*".

In addition to filling in missing information, DOMEs can be used to impose hypothetical management regimens, in order to simulate adaptation or other "what-if" scenarios. A "seasonal strategy" DOME allows a researcher to replicate management rules over multiple years for existing or climate change scenarios. The DOME associates metadata that describes the climate scenario, including global climate model, downscaling method, and represented historical or future time-slice, as well as identification of any adaptation strategies imposed. These metadata are an important component of maintaining a record of data provenance in the regional integrated assessment modeling process.

Table 1. Partial listing of DOME functions used to supply missing information to crop models or to impose hypothetical management regimens.

DOME function	Description
OFFSET	Add or subtract a number from the specified variable
OFFSET_DATE, DATE_OFFSET	Add or subtract a number of days from a specified date variable
MULTIPLY	Multiply the specified variable by a factor
PCTAWC	Compute water content in each soil layer from the specified percent available water
ICN_DIST	Initial soil nitrogen distribution function
AUTO_PDATE	Automatic planting date function
FERT_DIST	Fertilizer distribution function
OM_DIST	Organic matter application details
ROOT_DIST	Root distribution function
STABLEC	Stable soil carbon distribution
InitSW_dist	Initial soil water distribution
TAVAMP	Temperature average annual and amplitude
PADDY	Rice-paddy management function
AUTO_REPLICATE_EVENTS	Clones management events over multiple years
LYRSET	Soil-layer splitting function
AUTO_IDATE	Automatic irrigation date function
TRANSPOSE	Insert array value (e.g., for soil-layer data)

"Rotational strategy" DOMEs can be used to simulate long-term, continuous crop rotations. Like seasonal strategy analyses, the climate scenario and management regimen can represent historical, baseline, or hypothetical future conditions. Unlike seasonal strategy analyses, soil properties such as water and nitrogen content are initialized only once at the beginning of the continuous simulation period, rather than at the beginning of each planting season.

DOMEs operate through the specification of functions, which compute model inputs based on other available data. Table 1 presents a partial listing of DOME functions that were available as of this publication. All DOME functions are fully documented on the AgMIP research site (see http://research.agmip.org/display/it/Data+Interoperability).

Model-specific parameters that are not included in the ICASA-MVL can be supplied in a DOME, provided that the model translator is equipped to interpret the values. The DOME concept contributes to the vision of domain-specific languages for environmental modeling (Athanasiadis and Villa, 2013) by defining a set of domain-specific data operands for supplying missing information to crop models.

AgMIP crop model output (ACMO)

Simulated outputs from the crop models are harmonized and archived in the ACMO format to be used for further analysis, aggregation, and input to economic models. As with ACE, the ACMO schema conforms to the ICASA-MVL. Unlike ACE, the ACMO data are uniform, with each dataset containing exactly the same elements. Crop and economic modelers collaborated in the design to ensure that ACMO data are sufficient for use in the economic models, yet readily available from the outputs of most crop models. A full list of ACMO variables are listed and defined on the AgMIP research site (http://research.agmip.org/display/it/Data+Interoperability). Because of the consistent content of metadata and simulated outputs, regardless of model or simulation scenario, ACMO data are stored in a tabular text format.

Data translation libraries

Three types of translation tools were developed to support AgMIP crop modeling activities: (1) ACE input translators, which allow data to be imported from various formats into the ACE harmonization, (2) ACE output translators, which take combined ACE and DOME data and generate crop model ready files for crop simulation models, and (3) ACMO translators, which convert crop model outputs into a harmonized format. In most cases, these programs were developed as libraries of translation tools which can be implemented in multiple ways, including desktop applications, web services, and parallel-processing platforms for large-scale modeling applications. Table 2 lists the status of ACE input, ACE output, and ACMO translators completed or in development. The AgMIP regional research teams use the desktop applications discussed herein for use in regional integrated assessments crop modeling analyses.

ACE input translators: Input translators facilitate conversion of raw data from a variety of formats including spreadsheet templates encoded with ICASA variable names. Several spreadsheet templates have been provided to AgMIP users to accommodate differing levels of detail and widely varying management options, including templates for farm survey data, rice-paddy management systems, fertilizer trials, breeder trials, detailed crop experiments, and other options. The primary formatting requirement is that the data elements entered into a spreadsheet must be identified with an ICASA variable name and must conform to units specified in the ICASA-MVL. Examples of these spreadsheet templates are provided in the supplemental data at http://research.agmip.org/display/it/Data+Interoperability.

Model-specific input translators have been developed to import model data directly from the input files for some of the models actively used in AgMIP research. As of this publication, only the DSSAT and CropGrow-NAU ACE input translators

Table 2. Status of translator development for 13 participating models (Porter *et al.*, 2014). Translators are labeled "operational" if they have been implemented in translation applications; "nearing completion" if a team is actively working to finalize the translator; "under development" if work has begun, but is not yet complete; and "—" indicates that translator development has not begun. Reprinted from Poster *et al.* (2014), with permission from Elsevier.

Crop model	ACE input translator status	ACE output translator status	ACMO translator status	Reference for crop model
APSIM	Under development	Operational	Operational	Holzworth *et al.*, 2013; Keating *et al.*, 2003
AquaCrop	—	Nearing completion	Under development	Hsiao *et al.*, 2009; Raes *et al.*, 2009; Steduto *et al.*, 2009
CropGrow-NAU	Operational	Operational	—	Zhu *et al.*, 2004
CropSyst	Under development	Operational	Operational	Stöckle, *et al.*, 2003
DSSAT	Operational	Operational	Operational	Hoogenboom *et al.*, 2010; Jones *et al.*, 2003
EPIC	—	Under development	—	Izaurralde *et al.*, 2006
InfoCrop	—	Under development	—	Aggarwal *et al.*, 2006
ORYZA2000	—	Under development	—	Bouman and Van Laar, 2006; Bouman *et al.*, 2001
RZWQM2	—	Under development	—	Ahuja *et al.*, 2000
SALUS	—	Operational	—	Basso and Ritchie, 2012; Basso *et al.*, 2010
SarraH	—	Under development	—	Sultan *et al.*, 2013; Oettli *et al.*, 2011
STICS	Under development	Operational	Operational	Brisson *et al.*, 2009; Brisson *et al.*, 2003
WOFOST	—	Operational	Operational	Van Diepen *et al.*, 1989

were operational. Additional translators for APSIM, CropSyst, and STICS are currently under development. These model-specific input translators will allow many existing datasets associated with each modeling platform to be harmonized and made available to AgMIP researchers, in the formats specific to each model.

The AgMIP climate team developed a standard format file for disseminating daily weather records, as described in the AgMIP protocols (Rosenzweig *et al.*, 2011). Translators to convert these weather files to the ACE harmonized format were the first AgMIP input translators to be developed.

ACE output translators: Translators that convert data from the harmonized ACE format to model-ready formats were written for 13 models at the time of this publication (see Table 2). Five of the translators (APSIM, CropGrow-NAU, DSSAT, STICS, WOFOST) are used in a desktop translation utility and are currently available for use by AgMIP researchers. Additionally, the CropGrow-NAU, CropSyst, and EPIC models have translators linked to their respective model or user interfaces, which thus allows these models to use the ACE data format directly. Other ACE output translators are in various stages of development and will be available for future AgMIP modeling activities.

ACMO translators: ACMO translators are also currently available to AgMIP researchers for the APSIM, CropGrow-NAU, CropSyst, DSSAT, and WOFOST models. Translators for other participating models listed in Table 2, are being developed. The ACMO harmonized data include all metadata that describe the simulation, including links to the ACE and DOME data used in the simulation. These metadata are typically not available from the outputs of crop model simulations and so the ACMO file is partially created at the time of ACE data translation to crop model formats. After crop model simulations are complete, the ACMO translators combine the metadata with the simulated outputs into the ACMO harmonized file, in comma-delimited (csv) format.

Data provenance

Reproducibility and integrity of archived data is extremely important for AgMIP researchers. The data resulting from each analysis must include information related to the sources as well as to all intermediate modifications. The metadata included with ACE, DOME, and ACMO data enforce these provenance requirements. Metadata for each ACE dataset may include the source of field, weather, and soils records, the names and institutions of researchers involved in the collection of data, and location of measurements. DOME metadata include identification of the management regimen, adaptation scenario, and climate scenarios. Simulated model outputs in the ACMO file are connected via metadata to the raw data collected by researchers; the transformations and additions supplied by DOMEs; and the model name and version used for each simulation. The metadata fully identify the modeled scenario and provide the identification of simulated model outputs with the ACE datasets and DOMEs used to generate the model inputs.

Archived ACE, DOME, and ACMO datasets are tagged with hash codes which are generated from the data contents at the time of translation to harmonized format. This allows verification of the data integrity at every step in the processing chain. Hash codes are used to link the ACMO simulation outputs to the ACE and DOME data used to drive the simulation. When an existing archived ACE dataset is modified in a way that would affect simulated outputs for one or more models, a new unique hash code is generated and the new dataset is stored, linked to the original. In this way, a history of modifications can be maintained. Modified datasets are available to users, but a warning message is issued that the data may have been superseded.

Implementation of AgMIP Data Interoperability Tools

Desktop applications

All desktop data-translation applications described herein are available for download from the AgMIP toolshed at tools.agmip.org. These tools have been used to generate model inputs for ensemble modeling exercises for AgMIP regional integrated assessments, model intercomparison projects, and other crop modeling activities. All AgMIP software products, including the data translators and the desktop applications, have been developed as open-source projects and are available for download or contribution from github.com/agmip.

QuadUI

ACE input and output translators are bundled with the DOME interpreter in the QuadUI desktop application (see Fig. 3). The utility reads user-specified, site-based crop modeling data from one of several different formats and translates these data into model-ready input files for multiple models. QuadUI combines ACE and DOME data into a single JSON data-stream, which is supplied to the selected model-specific ACE output translators.

ACE data to be translated to crop model formats are provided by the user in a zip-format archive file containing one or more csv-format files containing the survey, soils, and weather data. These data were previously prepared using the standard data templates for ACE input translators, or other acceptable data input formats as described previously. Optionally, one or more field overlay DOME files can be provided in a single zip-format archive file. Seasonal strategy DOMEs, also optional, can be bundled together and specified by a user for generating multi-season simulations. Rotational strategies have not yet been implemented in QuadUI.

Model-ready files are written to a directory selected by the user, with separate sub-directories created for each crop model selected. A metadata file is also created at the time of translation for use in later generation of the ACMO files. QuadUI also generates compressed JSON files for ACE and DOME data. These compressed ace-binary files are the format that can be uploaded to an online AgMIP database.

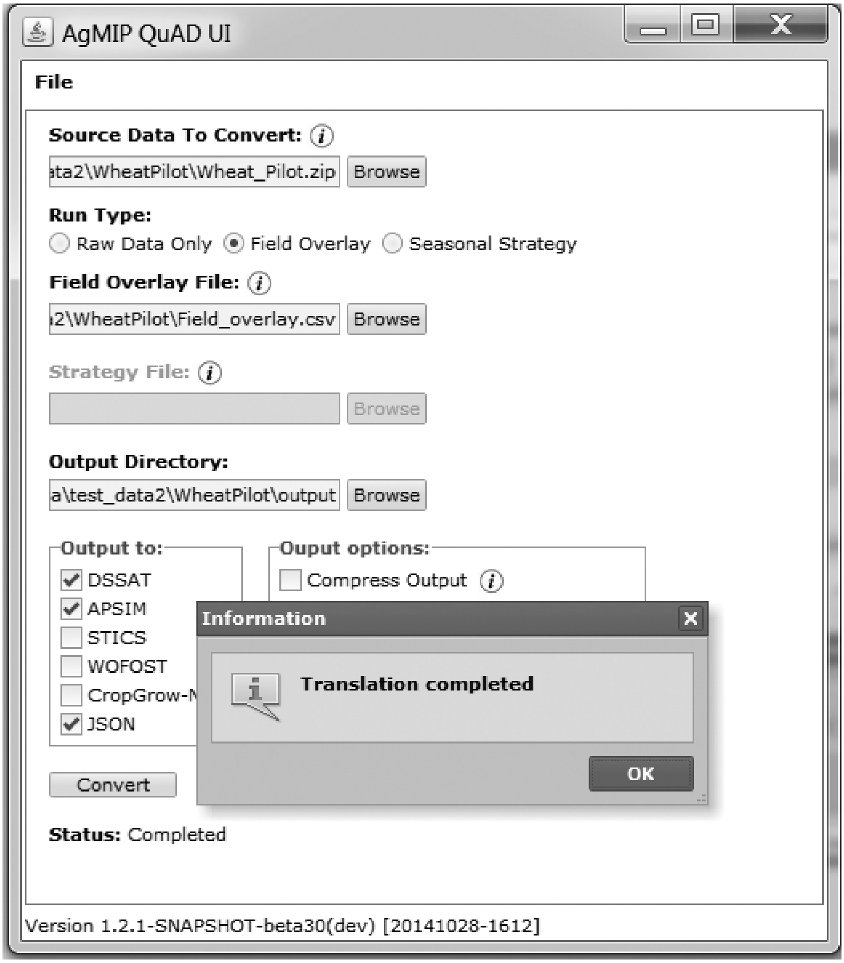

Fig. 3. QuadUI desktop application for conversion of raw data to model-ready formats. Users select the data file and optionally one or more DOME files. Data are converted into harmonized format and to one or more model-specific formats.

ADA

Raw data from surveys or field experiments are often entered into spreadsheets to facilitate data entry, conversion of units, and organization of data. The AgMIP Data Assistant (ADA, Fig. 4) desktop application exports multiple worksheets in a Microsoft Excel spreadsheet to csv format. Each worksheet is written to a separate csv file and the files are compressed in zip format.

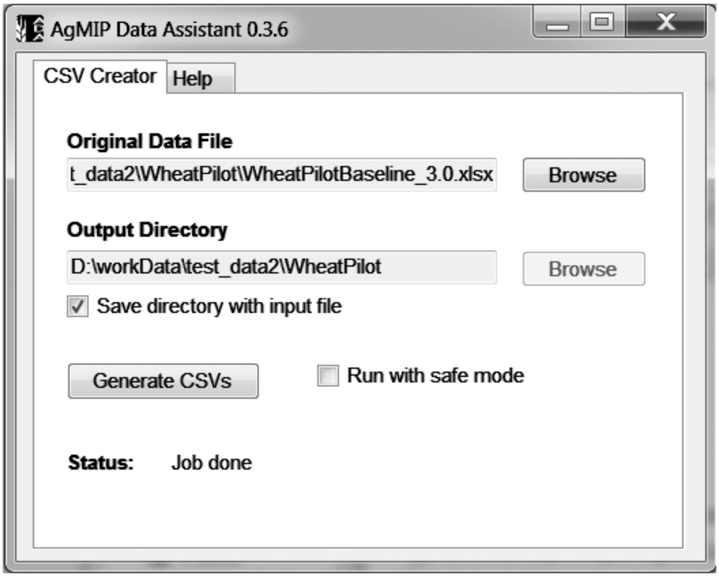

Fig. 4. ADA desktop application exports Excel worksheets to csv format in preparation for input to QuadUI.

ACMOUI

The ACMOUI application, shown in Fig. 5, is a separate desktop utility for harmonizing crop model outputs from multiple models. The model-specific ACMO translators append simulated outputs from the selected crop model to the metadata file generated by QuadUI. ACMO file names are generated from metadata and are of the form: "ACMO-Region-Stratum-Climate_ID-RAP_ID-Management_ID-Crop_model.csv". For example, an ACMO file from a DSSAT simulation of conditions in the Machakos region of Kenya for current climatology might result in a file name of "ACMO-Machakos-1-0XFX-0-0-DSSAT.csv", where OXFX is the climate identification for 1980–2010 baseline climatology using 2.5 minute WorldClim data and historical observations (see climate scenario naming conventions in Appendix 1, Part 1 in this volume).

AgMIP data interchange

The AgMIP Data Interchange (data.agmip.org) is the portal for accessing data relevant to AgMIP researchers. As of this publication, the site provides access to site-based datasets used in crop modeling activities in the AgMIP crop site database. These data include survey data associated with the RIAs, field research experimental data, and yield trial data from private industry and research institutions. These data include not only the field-observed data (ACE data), but also DOME (model

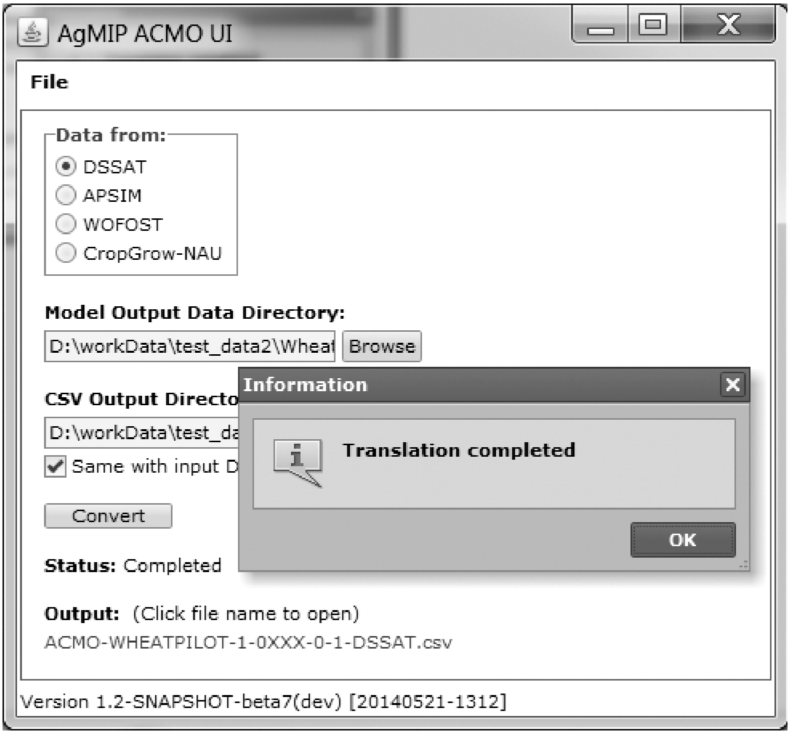

Fig. 5. ACMOUI converts simulated outputs from multiple crop models to harmonized ACMO format.

assumptions and scenarios) and ACMO data (simulated outputs) from crop modeling exercises. Data are uploaded, stored, and downloaded in the AgMIP harmonized data formats. Data translation utilities (QuadUI and ACMOUI) are used to generate the ACE- and ACMO-formatted data for upload to the site. Datasets that are downloaded from the site can be processed through QuadUI to generate crop model ready input files.

AgMIP protocols advocate open data policies and the intent of the AgMIP Data Interchange is to make these datasets freely available to anyone who wishes to use them. It is also recognized that researchers may wish to withhold publication of some datasets until their research results have been published. Some data used in AgMIP activities are not in the public domain, particularly historical weather observations which may be considered intellectual property by the entities that collected the data. For this reason, the AgMIP crop site database has three levels of data accessibility: (1) datasets which are freely available under the Creative Commons Attribution license (see creativecommons.org/licenses/by/3.0/us); (2) datasets that have been uploaded to the database, but that are restricted from download for a specified period of time before they are made available under the same Creative Commons Attribution license; and (3) datasets that are permanently restricted from download due to intellectual property issues. In the third case, only the metadata describing the data are stored in the AgMIP Data Interchange. Users who request these data are directed to the owner. Open datasets are available for download to anyone accessing the AgMIP Data Interchange; however only registered users have permission to upload data to the site.

The databases for ACE, DOME, and ACMO were designed using a Riak platform (basho.com/riak). Riak is an open-source, distributed database designed for scalable, fault-tolerant operation. The AgMIP implementation consists of multiple data nodes at facilities involved with AgMIP activities. As of this publication, the first data node was in testing phase at the University of Florida, with future nodes planned for University of Passo Fundo, Brazil; the International Crops Research Institute for the Semi-Arid Tropics (ICRISAT), India; and other locations.

AgMIP researchers are active in many countries and regions around the world. The data that are served by AgMIP must be made readily available in a format that can be searched and accessed efficiently by researchers who may lack access to high-speed internet connections. Many ensemble modeling approaches use 30 years or more of daily weather data associated with each climate scenario, climate model, downscaling method, and time-slice. Thus, weather datasets required for regional integrated assessments can be large. All data are uploaded and downloaded from the server use a compressed format (ace-binary) to minimize transfer latency.

In the future, additional data linkages to the AgMIP Data Interchange will allow access to climate datasets, soils databases, socio-economic model input and output data, and raster data relevant to gridded crop modeling applications.

Open Data Journal for Agricultural Research

AgMIP, in collaboration with MACSUR, Global Yield Gap Atlas, CIMSANS, and Wageningen UR, is developing a data journal for publishing peer-reviewed datasets associated with agricultural research. The *Open Data Journal for Agriculture Research* (*ODjAR*) will act as a central hub for storing, curating, and publishing the datasets as a resource for the future where publications and their authors get appropriate credit through citations and digital object identifiers for future reference. This valuable resource will allow datasets to be made available to researchers after a funding period ends and the research is published.

Next-generation AgMIP tools

The first-generation desktop utilities, QuadUI and ACMOUI, currently used in AgMIP RIAs, were developed quickly in response to an immediate need by RRTs.

It is recognized that the iterative process for generating and manipulating ACE and DOME data to produce successful and accurate simulations with multiple models, can be a cumbersome process. The list of variables in the ICASA-MVL is extensive and can be overwhelming for a user who is not familiar with the terms and structure of the data definitions. The format of the DOME is essentially a domain-specific language (DSL), which is often difficult for researchers to learn. For these reasons and others, we recognize that the next-generation user interfaces must facilitate data entry, manipulation and translation in framework that is easier to use.

The AgMIP IT team has begun conceptual design of an enhanced desktop utility for data handling that will combine the capabilities of QuadUI, ACMOUI, and the web interface for the AgMIP Data Interchange. Researchers will use this multi-function application for direct input of ACE and DOME data and to perform the process of refining the inputs to generate useful and accurate simulations. Models which can be run with a command line will be accessed directly from the interface, with ACMO outputs generated automatically. Direct linkage to the server will allow a user to search the database, download datasets, manipulate the data, and upload results to the server, all within a single desktop application.

Under a National Science Foundation project called Framework to Advance Climate, Economic and Impact Investigations with Information Technology (FACE-IT, www.faceit-portal.org), researchers are developing a web-based, interactive platform to support modeling activities for AgMIP and other research groups. This platform will allow a user to create complex workflows within a visual interface; connect the workflow to diverse data sources; visualize the data; and share the workflow, data, and outputs with other researchers conducting similar types of research. The platform uses the Galaxy workflow engine, used extensively in data-intensive biology applications, but under modification for use in earth science domains. For this NSF project, the AgMIP RIA process represents a use case. Workflow applications within the FACE-IT toolshed include the DSSAT and APSIM biophysical models; data translation utilities developed for AgMIP; climate data manipulation algorithms used by AgMIP climate scientists; and data visualization applications developed for FACE-IT. It is envisioned that this platform will facilitate ensemble modeling efforts for researchers worldwide as it moves into an operational phase in 2015.

Use of Data Translation Tools in the RIA Process

The AgMIP regional research teams in South Asia and Sub-Saharan Africa use AgMIP protocols to conduct regional integrated assessments to quantify the effects of climate change on food security in their regions. The teams use multiple climate, crop, and economic models to answer three key core questions:

1. What is the sensitivity of current agricultural production systems to climate change?
2. What is the impact of climate change on future agricultural production systems?
3. What are the benefits of climate change adaptations?

The AgMIP regional integrated assessment process requires collaboration among a multi-disciplinary team to provide consistent and cohesive inputs at each phase of the process for climate, crop, and economic analyses. These processes are described in detail in the AgMIP Regional Integrated Assessment Handbook (see Appendix 1, Part 1 in this volume) and throughout this volume. This use case focuses only on the crop modeling process, which uses outputs of climate models as input and which generates inputs for the economic modeling phase of the assessments. A simplified data-flow diagram for the crop modeling process is shown in Fig. 6, which shows the use of the three types of data translators and the two desktop utilities, QuadUI and ACMOUI.

Many sets of crop modeling simulations are required in order to evaluate current climate and technology conditions, future climate conditions with current technology trends and future climate conditions with adaptation. Future systems are simulated for multiple climate models, climate scenarios, and time-slices; using multiple crop models; and multiple site years. This results in hundreds of thousands of simulations which are to be evaluated and compared and used as input to regional economic models. The processes described below are for analysis of the historical conditions represented by the collected survey data. The data-translation process for analyses of seasonal strategies and future climate and adaptation scenarios is similar, but includes the application of additional seasonal strategy DOMEs.

The use of translation tools in the regional integrated assessments is part of an iterative process. The site-based data are collected, converted into the correct format and units as specified in the ICASA-MVL, and input to a survey data template, typically in a spreadsheet. The templates are modified by the crop modelers to include the appropriate ICASA variables associated with the available survey data for each regional analysis. Data are entered on separate worksheets for field and management data, soil data, and for each weather station.

A field overlay DOME is created using a spreadsheet template to supply data required by the crop models, but not provided in the farm or household survey data. DOME templates are available for users to modify with their site- and region-specific crop model inputs. Multiple DOMEs may be necessary to apply selective parameters to different soil types or socio-economic strata or to impose spatial variability among sites in an assessment.

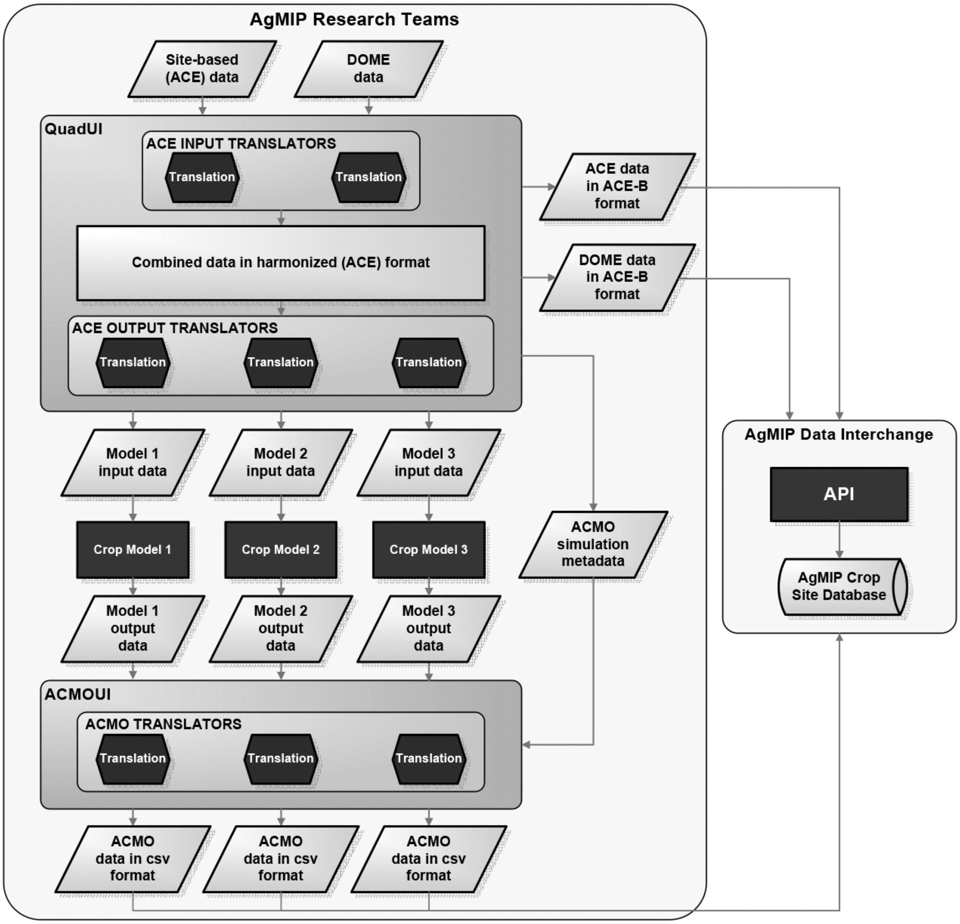

Fig. 6. Crop modeling data-flow diagram for AgMIP RIAs showing ACE input, ACE output, and ACMO translators as implemented in desktop utilities QuadUI and ACMOUI (Porter *et al.*, 2014). Reprinted from Porter *et al.* (2014), with permission from Elsevier.

ACE and DOME data are converted from spreadsheet format into csv format, then converted into a compressed zip format, either manually or using the AgMIP data assistant (ADA) desktop utility.

Generation of ACE and DOME data that produce error-free simulations for multiple models and multiple sites is an iterative process. QuadUI is used to read the survey and DOME data from the zipped archives, combine them into a single data stream, and translate into user-selected model formats. Syntax errors, misspelling of key variable names or DOME functions, or omission of critical information can all cause the translation to fail. When translations occur without error, the models may still fail if required model inputs are omitted. The models may run without error but produce results that do not adequately describe actual regional yields or other observed data. At each stage, the user must troubleshoot the problems using the QuadUI log, model outputs, and error messages, and by revising the survey and DOME sheets as needed to correct problems.

Once the simulations for historical conditions are successful, ACMOUI is used to generate the ACMO harmonized model output data for each analysis and each crop model. These data are then available in csv format for use by the regional economic models.

At the time that data are translated to model-ready formats, QuadUI generates a file that contains metadata for the simulation. This ACMO_meta.dat file is read by ACMOUI and combined with the simulated crop model outputs to generate the final ACMO file for each simulation. Thus, the final product from the process, the ACMO file, contains sufficient information regarding the scenario modeled to allow a researcher to recreate the simulations.

Analyses of multi-season systems for current and future climate conditions and adaptation scenarios follow the same process, and use the same field-overlay DOMEs. Seasonal strategy DOMEs are applied to impose automatic planting date rules, future management regimens, and to simulate over 30 weather years (current or future).

The final step in the process is for the user to upload the ACE, DOME, and ACMO data to the AgMIP crop site database using the AgMIP Data Interchange web interface (data.agmip.org/). ACE and DOME files are uploaded in compressed ACE (ace-binary) format; ACMO files are uploaded in csv format.

Discussion and Conclusions

The AgMIP harmonized data solution was developed in response to a need by the AgMIP community of researchers for interoperability of data for multiple crop models for assessment of climate change on food security. The ACE harmonized data format provides a simple, flexible, and extensible means of handling site-based agricultural data from diverse sources and for making them available as input to multiple crop models. The design of the data structures and modifications to the underlying ICASA data dictionary were collaborative efforts between the 13 crop modeling groups that participated in AgMIP Software Development Sprints in 2012 and 2013 (see research.agmip.org/display/itwiki/AgMIP+Development+Sprints) and the economists who will use outputs from the crop models.

The key-value schema of the ACE harmonized data format provides an efficient and flexible means of defining and archiving the irregular site-specific data from diverse sources. The DOME provides a means for users to supply information such as assumed management details, initial conditions, simulation time-span, or hypothetical management regimens, consistently and transparently, to multiple models. Harmonization of simulated model outputs facilitates consistent analysis and interoperability of the results from multiple models and makes the data more readily available for input to economic models.

The expertise of the modeler is critical in the RIA process. The tools and data provide only the input files for running the model, after which the modeler must step in, and check for missing values, out-of-bounds parameters, syntax errors, or inaccurate inputs, then correct these through iterative passes of DOME inputs. The modeler remains responsible for calibrating the model, validating the outputs, and generating appropriate scenarios for different climate, technology, or variety specifications. This focus on the practices of the modeler has an important implication in that the tools must be flexible enough to handle both the data used in the experiment and the model running the experimental data. This flexibility comes at a cost because for each combination of an experimental dataset and a modeling tool, an intervention of the modeler is required to fill parameters and to validate the model, including considering whether a given experiment can be run with a given model, and which experimental data can be adequately included. For example, it does not make sense to run an experimental dataset with detailed observations of nutrient dynamics through a crop model that lacks modules to compute nutrient dynamics.

The current implementation of AgMIP data interoperability tools uses spreadsheet templates and desktop utilities to allow researchers to harmonize and translate site-based data. These tools filled an immediate need for the researchers but remain difficult to use. Next-generation tools will include more interactive user interfaces and web-based approaches. The existing data schemas and translators are sufficiently flexible to be implemented in next-generation utilities without redesign.

The AgMIP data management system was designed to address the needs of a large, distributed, multi-model research community by adopting a rich, extensible vocabulary to describe site-based agricultural data from diverse sources; flexible data structures that allow both detailed and lesser-quality records to be stored in an efficient database; and the ability to provide a consistent set of assumptions and parameters which can be extrapolated to multiple models. For the first time ever, researchers and modelers are able to use these tools to run an ensemble of models on a single, harmonized dataset. This allows them to compare models, leading ultimately to model improvements.

Perhaps the most important outcome is that the AgMIP project has provided a platform that facilitates researcher collaboration from many organizations, across many countries. The level of cooperation between these groups is unprecedented and has already resulted in data interoperability tools that will benefit the large crop modeling community of researchers by making data from a wide variety of sources available to any model user. This would have been very difficult to achieve without the AgMIP data standards described in this chapter. As AgMIP grows to include more regions, researchers, and modeling groups, the harmonized data format and translation tools will continue to be valuable resources.

Acknowledgements

The authors acknowledge the efforts of numerous model experts and software developers who participated in code sprints and otherwise contributed to developing, testing, and implementing ACE, DOME, and ACMO data schemas. These contributors are listed in the following table.

Crop model	Translator developers
APSIM	D. Holzworth (CSIRO, Australia), I. Athanasiadis (Democritus University of Thrace, Greece)
AquaCrop	D. Raes (KU Leuven University, Belgium), R. Knapen (Alterra WUR, the Netherlands)
CropGrow-NAU	Z. Yan, L. Liu, X. Qiu (Nanjing Agricultural University, China)
CropSyst	R. Nelson (Washington State University, USA)
DSSAT	C. Porter, M. Zhang (University of Florida, USA)
EPIC	R. Sahajpal (University of Maryland, USA)
InfoCrop	N. Kalra (Tata Chemicals, India), S. Pamuru (Concat Systems, Inc., USA)
ORYZA2000	E. Manalo (International Rice Research Institute, Philippines), B. Sailaja (Directorate of Rice Research, India)
RZWQM2	P. Bartling (USDA ARS, USA)
SALUS	B. Baer (Michigan State University, USA)
SarraH	H. Songoti (Agrhymet, Niger), M. Zhang (University of Florida, USA)
STICS	D. Ripoche, J. Cufi (INRA, France)
WOFOST	J. teRoller, D. van Kraalingen, S. Janssen (Alterra WUR, the Netherlands)

We also recognize the larger AgMIP community including the members of the AgMIP regional research teams in South Asia and Sub-Saharan Africa, whose suggestions guided the development of the software products. Much of the support for development of the ACE data format, the data translation tools and the desktop and web-based user interfaces was provided under a grant by the UK Department for International Development (DFID).

References

Aggarwal, P. K., Kalra, N., Chander, S., and Pathak, H. (2006). InfoCrop: A dynamic simulation model for the assessment of crop yields, losses due to pests, and environmental impact of agro-ecosystems in tropical environments. I. Model description, *Agric. Syst.*, **89**(1), 1–25.

Ahuja, L. R., Rojas, K. W., Hanson, J. D., Shaffer, M. J., and Ma, L. (eds.) (2000). *The Root Zone Water Quality Model*, Water Resources Publications LLC, Highlands Ranch, CO, p. 372.

Athanasiadis, I. N. and Villa, F. (2013). A roadmap to domain specific programming languages for environmental modeling: Key requirements and concepts, *Proc. DSM-13*, ACM, pp. 27–32.

Basso, B., Cammarano, D., Troccoli, A., Chen, D., and Ritchie, J. T. (2010). Long-term wheat response to nitrogen in a rainfed Mediterranean environment: Field data and simulation analysis, *Eur. J. Agron.*, **33**, 132–138.

Basso, B. and Ritchie, J. T. (2012). Assessing the impact of management strategies on water use efficiency using soil-plant-atmosphere models, *Vadose Zone J.*, **11**(3).

Batjes, N. H. (2009). Harmonized soil profile data for applications at global and continental scales: Updates to the WISE database, *Soil Use Manage.*, **25**, 124–127.

Boote, K. J., Jones, J. W., White, J. W., Asseng, S., and Lizaso, J. I. (2013). Putting mechanisms into crop production models, *Plant, Cell, Environ.*, **36**, 1658–1672.

Bouman, B. A. M., Kropff, M. J., Tuong, T. P., Wopereis, M. C. S., Ten Berge, H. F. M., and Van Laar, H. H. (2001). *ORYZA2000: modeling lowland rice*, International Rice Research Institute, Los Baños, Philippines, and Wageningen University and Research Centre, Wageningen, the Netherlands, p. 235.

Bouman, B. A. M. and Van Laar, H. H. (2006). Description and evaluation of the rice growth model ORYZA2000 under nitrogen-limited conditions, *Agric. Syst.*, **87**, 249–273.

Brisson, N., Gary, C., Justes, E., Roche, R., Mary, B., Ripoche, D., Zimmer, D., Sierra, J., Bertuzzi, P., Burger, P., Bussiere, F., Cabidoche, Y. M., Cellier, P., Debaeke, P., Gaudillere, J. P., Henault, C., Maraux, F., Seguin, B., and Sinoquet, H. (2003). An overview of the crop model STICS, *Eur. J. Agron.*, **18**, 309–332.

Brisson, N., Launay, M., Mary, B., and Beaudoin, N. (2009), *Conceptual basis, formalisations and parameterization of the STICS crop model*, Editions QUAE, Versailles, France.

FAO/IIASA/ISRIC/ISSCAS/JRC (2012). *Harmonized World Soil Database (Version 1.2)*, FAO, Rome, Italy and IIASA, Laxenburg, Austria. Available at: http://webarchive.iiasa.ac.at/Research/LUC/External-World-soil-database/HTML/. Accessed on 7 March 2014.

Holzworth, D., Huth, N., deVoil, P., Zurcher, E., Herrmann, N., McLean, G., Chenu, K., van Oosterom, E., Snow, V., Murphy, C., Moore, A., Brown, H., Whish, J., Verrall, S., Fainges, J., Bell, L., Peake, A., Poulton, P., Hochman, Z., Thorburn, P., Gaydon, D., Dalgliesh, N., Rodriguez, D., Cox, H., Chapman, S., Doherty, A., Teixeira, E., Sharp, J., Cichota, R., Vogeler, I., Li, F., Wang, E., Hammer, G., Robertson, M., Dimes, J., Carberry, P., Hargreaves, J., MacLeod, N. C. M., Harsdorf, J., Wedgewood, S., and Keating, B. (2013). APSIM — Evolution towards a new generation of agricultural systems simulation, *Environ. Model. Software: Agric. Syst. Model. Software*. Available at: http://www.sciencedirect.com/science/article/pii/S1364815214002102. Accessed on 10 October 2014.

Hoogenboom, G., Jones, J. W., Wilkens, P. W., Porter, C. H., Boote, K. J., Hunt, L. A., Singh, U., Lisazo, J. I., White, J. W., Uryasev, O., Royce, F. S., Ogoshi, R., Gijsman, A. J., and Tsuji, G. Y. (2010). *Decision Support System for Agrotechnology Transfer Version 4.5* [CD-ROM], University of Hawaii, Honolulu, Hawaii.

Hsiao, T. C., Heng, L., Steduto, P., Rojas-Lara, B., Raes, D., and Fereres, E. (2009). AquaCrop-The FAO crop model to simulate yield response to water: III. Parameterization and testing for maize, *Agron. J.*, **101**(3), 448–459.

Hunt, L. A., White, J. W., and Hoogenboom, G. (2001). Agronomic data: advances in documentation and protocols for exchange and use, *Agric. Syst.*, **70**, 477–492.

ICRISAT (International Crops Research Institute for the Semi-Arid Tropics) (1984). *Proc. Int. Symp. Minimum Datasets Agrotechnol. Transfer*, ICRISAT, India, Patancheru, India.

Izaurralde, R. C., Williams, J. R., McGill, W. B., Rosenberg, N. J., and Jakas, M. C. (2006). Simulating soil C dynamics with EPIC: Model description and testing against long-term data, *Ecol. Model.*, **192**(3), 362–384.

Jones, J. W., Hoogenboom, G., Porter, C. H., Boote, K. J., Batchelor, W. D., Hunt, L. A., Wilkens, P. W., Singh, U., Gijsman, A. J., and Ritchie, J. T. (2003). The DSSAT cropping system model, *Eur. J. Agron.*, **18**(3–4), 235–265.

Keating, B. A., Carberry, P. S., Hammer, G. L., Probert, M. E., Robertson M. J., Holzworth, D., Huth, N. I., Hargreaves, J. N. G., Meinke, H., Hochman, Z., McLean, G., Verburg, K., Snow, V., Dimes, J. P., Silburne, M., Wang, E., Brown, S., Bristow, K. L., Asseng, S., Chapman, S., McCown, R. L., Freebairn, D. M., and Smith, C. J. (2003). An overview of APSIM, a model designed for farming systems simulation, *Eur. J. Agron.*, **18**, 267–288.

NIST (2012). *Secure Hash Standard (SHS)*, National Institute of Standards and Technology, US Department of Commerce, FIPS Pub #180-4, Gaithersburg, MD, USA. Available at: http://csrc.nist.gov/publications/fips/fips180-4/fips-180-4.pdf. Accessed on 22 September 2014.

Oettli, P., Sultan, B., Baron, C., and Vrac, M. (2011). Are regional climate models relevant for crop yield prediction in West Africa, *Environ. Res. Lett.*, **6**, 014008.

Porter, C. H., Villalobos, C., Holzworth, D., Nelson, R., White, J. W., Athanasiadis, I. N., Janssen, S., Ripoche, D., Cufi, J., Raes, D., Zhang, M., Knapen, R., Sahajpal, R., Boote, K. J., and Jones, J. W. (2014). Harmonization and translation of crop modeling data to ensure interoperability, *Environ. Model. Software*, **62**, 495–508.

Raes, D., Steduto, P., Hsiao, T. C., and Fereres, E. (2009). AquaCrop-The FAO crop model to simulate yield response to water: II. Main algorithms and software description, *Agron. J.*, **101**(3), 438–447.

Rosenzweig, C., Jones, J. W., Hatfield, J. L., Ruane, A. C., Boote, K. J., Thorburn, P. J., Antle, J. M., Nelson, G. C., Porter, C. H., and Janssen, S. (2011). *AgMIP Protocols*. Available at: http://research.agmip.org/display/research/AgMIP+Protocols. Accessed on 22 September 2014.

Rosenzweig, C., Jones, J. W., Hatfield, J. L., Ruane, A. C., Boote, K. J., Thorburn, P. J., Antle, J. M., Nelson, G. C., Porter, C. H., Janssen, S., Asseng, S., Basso, B., Ewert, F., Wallach, D., Baigorria, G. A., and Winter, J. M. (2013a). The Agricultural Model Intercomparison and Improvement Project (AgMIP): Protocols and pilot studies, *Agric. Forest Meteorol.*, **170**, 166–182.

Steduto, P., Hsiao, T. C., Raes, D., and Fereres, E. (2009). AquaCrop-The FAO crop model to simulate yield response to water: I. Concepts and underlying principles, *Agron. J.*, **101**(3), 426–437.

Stöckle, C. O., Donatelli, M., and Nelson, R. (2003). CropSyst, a cropping systems simulation model, *Eur. J. Agron.*, **18**, 289–307.

Sultan B., Roudier, P., Quirion, P., Alhassane, A., Muller, B., Dingkuhn, M. Ciais, P., Guimberteau, M., Traore, S., and Baron, C. (2013). Assessing climate change impacts on sorghum and millet yields in the Sudanian and Sahelian savannas of West Africa, *Environ. Res. Lett.*, **8**, 014040.

Uehara, G. and Tsuji, G. Y. (1998). "Overview of IBSNAT" in Tsuji, G.Y., Hoogenboom, G., and Thornton, P. K. (eds.), *Understanding Options for Agricultural Production*, Kluwer Academic Publishers, Dordrecht, the Netherlands, pp. 1–7.

Van Diepen, C. A., Wolf, J., van Keulen, H., and Rappoldt, C. (1989). WOFOST: A simulation model of crop production, *Soil Use Manage.*, **5**(1), 16–24.

White, J. W., Hunt, L. A., Boote, K. J., Jones, J. W., Koo, J., Kim, S., Porter, C. H., Wilkens, P. W., and Hoogenboom, G., (2013). Integrated Description of Agricultural Field Experiments and Production: The ICASA Version 2.0 Data Standards, *Computers Electron. Agric.*, **96**, 1–12.

Zhu, Y., Cao, W. X., Shen, W. X., and Tian, Y. C. (2004). Dynamic knowledge model and decision support for rapeseed cultivation management, *Trans. Chin. Soc. Agric. Eng.*, **20**(6), 141–144 (in Chinese with English abstract).