Accepted Article

Rao, N H, **Big Data and Climate Smart Agriculture - Status and Implications for Agricultural Research and Innovation in India**, *Proceedings of the Indian National Science Academy*

DOI: 10.16943/ptinsa/2018/49342

Received date: 22.09.2017
Revised date: 01. 01.2018
Accepted date: 12.02.2018
Published Online: 13.02.2018

Big Data and Climate Smart Agriculture - Status and Implications for Agricultural Research and Innovation in India

N H Rao

KL Rao Chair Professor in Geospatial Sciences, Centre for Earth and Space Studies, University of Hyderabad, Hyderabad, India
Email: nhrao1954@gmail.com

Abstract

Climate change will increase the vulnerability of agricultural production systems, unless scientists and farmers reorient their present approaches toward making them climate smart or climate resilient. The integration of recent developments in big data analytics and climate change science with agriculture can greatly accelerate agricultural research and innovation for climate smart agriculture (CSA). CSA refers to an integrated set of technologies and practices that *simultaneously* improve farm productivity and incomes, increase adaptive capacity to climate change effects, and reduce green house gas emissions from farming. It is a multi-stage, multi-objective, data-driven, and knowledge based approach to agriculture, with the farm as the most fundamental unit for both strategic and tactical decisions. This paper explores how big data analytics can accelerate research and innovation for CSA. Three levels at which big data can enhance farmer field level insights and actionable knowledge for the practice of CSA are identified: (i) developing a predictive capability to factor climate change effects to scales relevant to farming practice, (ii) speeding up plant breeding for higher productivity and climate resilience, and (iii) delivery of customized and prescriptive real-time farm knowledge for higher productivity, climate change adaptation and mitigation. The state-of-art on big data based approaches at each of the three levels is assessed. The paper also identifies the research and institutional challenges, and the way forward for leveraging big data in research and innovation aimed at climate smart agriculture in India.

Key words: Climate smart agriculture, big data, food security, innovation

Climate Smart Agriculture with Big Data - Review of Current Status and Implications for Agricultural Research and Innovation in India

1. Introduction

Climate change intensifies the challenge of future food security (Campbell et al., 2016). Rising average temperatures, more variable rainfall and increasing frequency of extreme events, resulting from anthropogenic climate change (Fischer and Knutti, 2015) will increase vulnerability of agricultural production systems, unless scientists, farmers and agribusiness reorient their present approaches to make them climate smart or climate resilient (World Bank, 2015). Agriculture also accounts for 19 to 29 percent of total greenhouse gas (GHG) emissions that contribute to climate change, and the largest share of non-CO_2 GHGs. The Paris Agreement of 2015 mandates nations to take urgent action to reduce GHG emissions to limit global warming below 2^0C (UNFCC, 2015).Climate smart agriculture (CSA) is an important part of the solution to climate change problem, and is necessarily the future way of agriculture.

Emerging big data technologies promise new levels of scientific discovery and innovative solutions to complex problems (National Science and Technology Council, 2016). They can be leveraged to address complex problems of addressing climate change and its impacts on agriculture (Faghmous and Kumar, 2014; World Bank, 2016).The transformative potential of combining big data from crop genomics, phenomics, climate, remote sensing, and individual farms for generating scientific, economic, social and environmental value in agriculture is underscored by Monsanto's recent (2013) acquisitions of Climate Corporation, Precision Planting, and several other data resource and analytics companies; similar acquisitions and partnerships among agribusiness multinationals; and emerging start-ups in agriculture that leverage data analytics (Gilpin, 2014; Bomgardner, 2016).The Digital India initiative has also made India a fertile ground for diverse groups of scientists, students, analysts, businesses, and entrepreneurs to leverage big data for farmer and business value. Indian agriculture start-ups attracted over US $500 million investment in 2015, the third largest globally (after US and Israel), the second largest (after US) in the number of start-up ventures financed, and the largest globally in drones, irrigation technologies and data driven agricultural decision support systems (AgFunder, 2016).

The purpose of this paper is to assess the state-of-art on how big data tools and methods can be leveraged to integrate climate, crop and agricultural informatics with design and management of agricultural systems at farm level for climate smart agriculture. The potential opportunities, challenges, and the way forward for research and innovation aimed at climate smart agriculture in India are identified. The paper is structured as follows:

1. Big data - a perspective on technologies and potential in agriculture
2. Big data and climate smart agriculture, and
3. Roadmap for leveraging big data for climate smart agriculture in India

2. Big data - a perspective on technologies and potential in agriculture

Big data broadly refers to large, diverse, complex, longitudinal or distributed datasets generated from a variety of sources (instruments, sensors, internet transactions, email, video, click streams, and/or all other digital sources) available today and in the future (National Science Foundation, 2012). Traditional data analysis can also involve many observations, but it includes only a few variables to explain a phenomena. Big data changes this paradigm by recognizing that when data and data sources are large and diverse, they hold a lot of detail that can explain complex phenomena (Halevy et al., 2009). The value of data increases manifold if it can be linked and integrated with other data, irrespective of their source or form. The promise of new levels of knowledge discovery and economic value has made big data analytics one of the key tools of the 21st century. By 2030, big data is

expected to provide the foundation for a global second economy which can potentially exceed the size of first economy (Arthur, 2011).

Big data analytics is essentially an outcome of developments across three major components of the digital revolution (Kuneet al., 2016): (i) new digital data sources, (ii) more computing power (faster processers and networks, massive storage, parallel processing, cloud computing), and (iii) higher level analytics (machine learning, deep learning, natural language processing, visualization). Together, they enable creation of novel value by leveraging massive, structured, and unstructured data to generate powerful insights into complex phenomena.(Unstructured data does not conform to a pre-defined schema and cannot be easily searched or processed in traditional database systems). Sensors, search engines and social media are examples of sources of unstructured data (text, documents, images, videos, etc.). That only 5% of all data in the world is structured data (Gandomi and Haider, 2015), underscores the wide scope and significance of big data technologies.

Large volume, variety, and velocity are three basic characteristics of big data. Volume refers to size of data, while variety encompasses multiple data sources, variables, formats and heterogeneity (structured/unstructured data). Velocity refers to the frequency at which data is acquired, which can vary from seconds to years. Data veracity (uncertainty), variability (inconsistency) and value are also often included as additional characteristics of big data. Such data are too large to be stored or processed on a single computer using traditional software and database architectures (tables, excel sheets, SQL databases). While the size/volume of big data gets popular attention, the heterogeneity of sources, formats and lack of structure present its most difficult challenges (Davenport, 2013). The key idea of big data therefore also includes novel methods used for data integration, storage, processing, visualization, and analyses. A definition which covers all these aspects is: Big Data is data of such large size and complexity (large number of variables and diversity of their sources, structures, frequencies, and scales), that they require new computer and data architectures, techniques, algorithms, and analytics to manage and extract value and hidden knowledge **(**adapted from Schönberger and Cukier, 2013).

Creating value from big data involves five distinct steps: (i) data acquisition and storage, (ii) information extraction and cleaning, (iii) data integration, (iv) modeling and analysis, and (v) interpretation and deployment (Jagdish, 2015). Specific big data technologies analyze textual, video and audio data and link them to other data. Similarly, to deal with high volume, variety and velocity aspects of big data, machine learning technologies are used to rapidly fit, optimize and predict data. Further, as big data are too large to store in any single central data base, technologies for parallel storage and processing among several computers are deployed for faster and more balanced output. Finally, the visualization tools of big data enable users to interact with underlying algorithms, assess and interpret outcomes of analysis, and communicate with stakeholders.

Hadoop, an open source software built in java programming language, is the most widely used big data technology for distribution of data and sub-problems for parallel storage and processing (Davenport, 2015). It is a highly scalable, integrated environment for dividing, storing and processing both structured and unstructured data across multiple processors (nodes). Several versions of Hadoop are available from different vendors. Another commonly used tool is MapReduce, a programming model developed by Google for reliable, scalable and distributed computing across a group of linked computer nodes or processors. A recent version of the Hadoop-MapReduce framework (Apache Hadoop Release 3.0, December 2017) can be downloaded from http://hadoop.apache.org/releases.html.

Hadoop's java based file system (Hadoop File System or HDFS) stores both structured and unstructured data in small replicated blocks of uniform size (default 64MB). HDFS manages large data by splitting and storing data from each file in multiple files

distributed across blocks at multiple nodes. The file system uses a tree structure to store and identify data in two types of nodes, Namenode and Datanode. For each source of big data, Namenode identifies the source file name and its metadata. The data are stored in Datanodes, in blocks of distributed clusters of uniform storage capacity, and linked to their corresponding Namenode. This permits identifying, structuring, storing, querying, and processing the split data from each node independently. Hadoop's client interface enables reading and writing data to different files. MapReduce provides the tools for parallel processing of data in HDFS. Processing can occur with data stored either in HDFS (for unstructured data) or in a database (for structured data). The advantage is that MapReduce can process data at the node where the data file is located without transmitting it to an independent processor.

The Hadoop ecosystem has evolved in recent years to a platform that includes many different tools in addition to MapReduce. These include Mahout (a scalable machine learning and data mining library), Pig (a high-level data-flow language and execution framework for parallel computation), Spark (fast and general programming models for Hadoop data querying and processing applications and visualizations), and security and data management tools. Other programming languages, database systems and hardware architectures are also being added. Recent trends point to hybrid architectures that integrate Hadoop, databases, cloud sources (Assuncaoet al. 2015), electrical and optical networks (Rehman and Esmailpur, 2016), massive parallel processing technologies, and incorporate both established database and new approaches into a common platform. A typical schematic of big data storage and processing environment is shown in Fig 1. Many softwares in Hadoop platform are open source, but they require high degree of computer programming and analytics skills.

A key difference between big data and traditional data analysis is that the latter is hypothesis driven while big data analytics relies on machine learning to arrive at best fit models. The central feature of machine learning is its essentially 'theory-free' or hypothesis free' approach with focus on learning from data. This can result in challenges of interpretation, spurious correlations, and model fits.

Businesses were among the first to gain from the theory free approach of big data to predict consumer preferences from buying data, leading to new marketing strategies and higher profits (Davenport, 2013). But this should not be equated with improved scientific understanding of buyer behaviour, as marginal increases in targeted offers to consumers can translate to significantly higher profits. Other situations may not provide similar value with the theory free approach. A classic example of limitations of theory-free big data analytics (machine learning) is the Google Flu Trends (GFT) Model to track and predict flu outbreaks in USA (Ginsberg et al., 2009), and its later inadequacies (Butler, 2013). GFT estimated weekly influenza activity with a one day reporting lag, much shorter than US Centre for Disease Control's two week lag. But the theory-free big data analysis proved fragile as its success provoked internet searches by people who were healthy, leading to bias in data. From 2015, Google stopped publishing flu trend data and passed them to specialized organizations.

What the GFT model highlights despite its inadequacies is that machine learning can provide useful insights for domain practitioners to ask and resolve high level questions that might not otherwise be asked. Complementary domain knowledge can add significant value to machine learning outputs by providing better insights into identifying and testing more meaningful hypotheses for causal inferences (Maciejewskiet al., 2016; Dhar, 2013; Shriffin, 2016). This difference from essentially data driven business analytics is most critical for applications of big data analytics in scientific knowledge discovery domains such as agriculture.

Perhaps no other area is so alluring for big data-based innovations than agriculture (Wolfertet al., 2017; Jackson, 2016; Mckinsey& Company, 2016; Gilpin, 2014; Sonka, 2016). Big data in agriculture comes from big data of crop genomics and phenomics, and lots

of small data arising from wide spatial and temporal variations in climate, crops, farmers, land, soil, water, infrastructure, markets, socio-economic conditions, GHG emissions, environmental and climate impacts, etc. Increasingly, weather and remote sensing satellites, climate models and forecasts, and more recently micro satellites, drones and field sensors, have added to the volume, velocity and variety of agricultural data. Farmers too have demonstrated their capacities to engage with data driven information and advisories delivered on mobiles (Glendenning and Ficarelli, 2012). Their responses also constitute new data that augment agricultural big data. National initiatives like Digital India, which extends the digital network to 250000 village Panchayats, are expected to provide customized agricultural knowledge services to individual farmers. The Digital India Network is also a source of distributed big data that can be leveraged for tailored knowledge services to farmers, businesses, government and communities. But, for these initiatives to generate sustainable value, strengthening the interface between big data and domain, knowledge of CSA will be necessary.

3. Big Data and Climate Smart Agriculture

As agriculture constantly seeks new products, practices and technologies to enhance food security, and farmer and consumer welfare, the productive capacity of its natural resources base is shrinking. Climate change compounds the exigencies of food security and natural resource sustainability by exposing farming to greater uncertainties and risks of extreme events (WMO, 2010, Campbell et al., 2016).Agriculture also contributes to climate change with the major share (37%) of N_2O and CH_4 emissions (Paustian, 2016). Improved soil and crop management can substantially reduce emissions and restrict global temperature increases.

The multi-dimensional aspects of agricultural production under climate change are captured in FAO's definition of climate smart agriculture (CSA)as: "agriculture that sustainably increases productivity, resilience (adaptation), reduces/removes GHGs (mitigation), and enhances achievement of national food security and development goals."(FAO 2010). By this definition, CSA has three *concurrent* objectives: (i) sustainably increasing farm productivity and income, (ii) increasing adaptive capacity to climate change, and (iii) reducing GHG emissions. The fundamental land unit for operationalizing CSA is necessarily the farm.CSA is about proactively and precisely responding to more variable and extreme conditions under climate change by integrating strategies to improve crop planning, adaptation and agronomic management on the farm. Geographic and temporal specificity of information on climate change effects, risks, agronomic practices, and impacts at farm level is therefore critical for implementing CSA.

A number of strategies and farm technologies have emerged for practice of CSA (Rosenstocket al., 2016). These include breeding crops for resilience to stresses induced by climate change, shifting crop production to new seasons/regions, insurance against risks, and crop, soil and water technologies and practices to improve productivity, input efficiencies, carbon sequestration, and reduction of GHG emissions (ICF International, 2016). These responses present their own challenges. Plant breeding takes time and is limited by genetic variation within or across crops and environments. Generation of technologies for soil, water and crop management by traditional field experimentation in research station plots across multiple environments is also cumbersome and time consuming. The experiments and analysis focused on interaction between crop and macro-environment, have generally excluded farmers and management practices on their fields as variables. As a result, advice on implementation of CSA technologies and practices is generic, arbitrary, inconsistent, imprecise, slow, and lacking in scale and sustainability. The primary reason for this is lack of geographic and temporal specificity of climate change data and process-driven actionable

knowledge required for CSA. Similarly, market adaptive systems for climate change like agricultural insurance can be effective only if expected risks at farm level can be assessed with greater spatial specificity.

Big data's capacity for inclusion of heterogeneity - across farms, farmers, climates, crops, soils, natural resources, models, management strategies and outcomes, post production value chain systems, and other economic variables of interest - can boost geographic specificity, timeliness and scalability of actionable knowledge for CSA. This amalgamation of agriculture and big data is being seen as potentially the greatest accelerator of food production technology since the Green Revolution (World Economic Forum, 2016; Bomgardner, 2016).

For big data analytics to enhance research and innovation for strategic and tactical decision-making with the geographic specificity needed for CSA, it needs to be applied and integrated across three levels: (i) developing a predictive capability to factor climate change effects to resolutions compatible with farm planning, practice, and risk assessment, (ii) speeding up and improving precision of plant breeding for climate resilience, and (iii) providing farm level customized knowledge support for better crop planning, higher productivities, climate change adaptation and GHG reduction. The state-of-art of big data analytics at these three levels is reviewed.

(i) ***Developing predictive capability to factor climate change effects to farm level***: To support both strategic and tactical decisions for CSA, apriori assessments of climate change at farming scale precision are required at two levels: (i) projections, and (ii) predictions. Projections address the long-term (30, 50, 100 years) and provide information on intra-seasonal risks of extreme events under climate change (extreme events are events with a probability of occurrence of a climate variable above (or below) a threshold value near the upper (or lower) ends of the range of the observed values of the variable - IPCC 2012). Projections are valuable for strategic planning for CSA, eg. for germplasm screening, setting plant breeding goals, crop planning, designing water systems, etc. Climate predictions focus on short-range climate predictability (hourly, daily, weekly) under climate change to support climate smart agronomic management and operations on the farm.

Global climate models (GCMs) project future long term average changes and extremes of temperature, precipitation, and other variables for standard future GHG emissions scenarios characterized by IPCC as Representative Concentration Pathways (RCPs 2.6, 4.5 and 8.5) . The models discretize the earth's surface into thousands of grid cells at resolutions of 60 to 300 km (typical IPCC's CMIP5 set of GCMs are ~ 100 x 100 km). The coarse spatial resolution of GCMs does not allow calculating the climate variables at farming compatible spatial resolutions. Also, traditional agricultural research spans two to five year time horizons. But, CSA alters the time horizon to next 30-40 years or more, to factor in relationships between global change and local climate change, and risks of extremes (higher intensity and frequency of extreme events like heat shocks, high rainfall events, droughts, floods, etc.).CSA therefore challenges both GCMs and traditional agricultural research.

Currently available GCMs predict quite well the large-scale climate features such as circulation patterns, El Niño Southern Oscillation (ENSO), global mean temperature, and precipitation (to lesser extent than temperature).Statistical downscaling methods or regional climate models (RCMs) are used to merge information from GCMs with regional and local meteorological records and local geography to transform coarse-scale GCM simulations to climate information for smaller spatial units (Benestad, 2016). Downscaling assumes a systematic link between conditions taking place on a global scale and local conditions. The end result is a more spatially detailed picture of what future climate change could mean at local levels. The management and manipulation of dozens of climate models and their daily

outputs for 50-100 years, and of long term data from multiple sources for downscaling, is clearly a big data problem.

Downscaling global models to national/sub-continental scales and 50/25 km grids has been common in the past two decades, with projections up to years 2050-2100 for different emissions scenarios (Benestad,2016; Girvetzet al.,2013). Further, at these scales , machine learning algorithms like Support Vector Machine (SVM), Random Forest (RF), Artificial Neural Networks (ANN) or their hybrids have also been shown to be superior to traditional regression based climate downscaling (Tripathi et al,, 2006; Goly, 2014, Anandhi et al., 2008).Downscaling climate to higher resolutions (1km) and hyperlocal resolutions (farm, watershed) for agricultural or hydrologic applications has proved more challenging (Fig 2). Most attempts have involved downscaling from GCMs/RCMs by interpolation (egWorldClim 1km resolution climate surfaces for current and projected climate to 2070,http://www.worldclim. org/version1). The advent of big data with its tools of machine learning and deep learning have increased the predictive capacity of climate models at hyper local resolutions by systematically exploring the links between GCM/RCM outputs and high resolution local data (Vandal et al., 2017a, b; Ford et al., 2016; Bell et al., 2016; Thrasher et al., 2013).

Examples of big data based downscaling of climate data to finer resolutions include: (i) the NASA Earth Exchange (NEX) which provides downscaled CMIP5 climate models projections at 30 arc-second (~ 800m) resolution and daily time intervals for distribution through its portal (https://portal.nccs.nasa.gov/portal home/published/NEX.html; Thrasher et al. 2013); (ii) The Climate Corporation's seamless integration and extraction of public data from climate models, soil surveys, weather stations, weather radar, 10m x 10m soil maps, crop models, and farm maps on Google Earth to generate hyper-local weather forecasts for agricultural insurance and agri-advisory services field by field to US farmers (Bell et al. 2016), and (iii) IBM's Deep Thunder which incorporates a variety of inputs from satellites; climate model ensembles, terrestrial data (topography, soils, land use, vegetation, and water temperature) gathered by sensors aboard NASA spacecraft, U.S. Geological Survey and many private weather stations, to forecast the weather every 10 minutes, for each 1.5 square kilometer of farmland (Knowledge@Wharton, 2014). However, work in India on climate downscaling (Chaturvedi et al., 2012) has largely been limited to 25/50km resolution outputs from IMD's RCMs for climate change projections at these scales for the future emissions described by IPCC's scenarios, including the more recent emission pathways (RCPs).

The state-of-art in big data climate analytics is that downscaled information at finer resolutions relevant to CSA can be derived from GCMs and local data, and accessed by practitioners from private and public providers (Wang et al., 2016; https://nex.nasa.govhttps://nex;http://www.climsystems.com). The data can be in different forms (maps, time series, summaries etc.) and downloaded through web services and to GIS platforms. For the practice of CSA, a critical concern is evaluating and authenticating the various datasets for their applicability at local conditions before using them. A central question for CSA is also that of assessing uncertainty and risk from downscaled projections. Current practice is to use ensembles of climate projections with the algorithms (Ekstromet al., 2015) to enlarge the scope and range of data by including different sources of uncertainty (in emissions data, models, scenarios, downscaling methods). The challenge for CSA is to leverage finer resolution climate projections to generate better insights into plant breeding for better precision in choice of climate resilient crops/varieties, and for more efficient, effective and risk averse farm operations.

(ii) ***Speeding up and improving precision of plant breeding for climate resilience***: Plant breeding provides the primary genetic resources for farm scale climate stress resilience. But, conventional plant breeding cycles are long (5-20 years), designed for average macro-

environment conditions, and limited by existing gene pools. This limits the capacity to respond urgently and precisely to rapidly variable conditions and extremes under climate change. Even transgenic and molecular breeding approaches are relatively slow and uncertain as they depend on random integration of new gene sequences into plant genomes. The former are further slowed by regulatory systems for GMOs. Further, plant breeding target genes for CSA are most likely to be tolerance to abiotic stresses and general productivity, adaptation and mitigation improvements (increasing photosynthesis, altering flowering times and root systems; increasing nutrient uptake from applied fertilizers, etc.). These are multigenic traits, and thus difficult and slow with conventional breeding. Combining big data from genomics, phenomics, and climate models with new genetic engineering tools can potentially speed up and improve geographic specificity and precision of plant breeding.

The exponential increase in genome and expressed sequence data of thousands of varieties of agriculturally important crops in the past two decades (Koleet al. 2016; Edwards et al. 2016; Varshney, 2016) has made genomics a big data science (Stephens et al. 2015). A similar explosion in high throughput plant phenotyping (HTPP) data for a wide range of crops and environments is occurring as a result of advances in automated sensing and imaging technologies. Storing and analyzing large genomic, HTPP and environmental data for more efficient dissection of genes corresponding to various biotic and abiotic stress-related adaptive traits in different environments is a big data problem. Big data tools of machine learning can more effectively analyze HTPP (Navarro et al., 2016) and genomics data (Singh et al. 2016). Integrated analysis of phenomics and genomics data with machine learning tools can help plant breeders analyze much larger number of attributes, identify novel candidate genes, automate phenotypic ratings and generally improve the genomic prediction accuracy (Zhang et al., 2017).

Simultaneous advances in new molecular biology techniques like genome editing, permit faster, more controlled and precise integration of new genes than the earlier plant breeding and transformation methods (Scheben and Edwards, 2017; Altpeteret al., 2016). A key technical advance in genome editing is the CRISPR tool which has vastly enabled rapid innovation in DNA modification by improving precision, increasing reliability, shortening time and reducing costs. A major advantage of CRISPR is, though genetic information of the cutting protein is foreign DNA, it can be completely removed after modification of the genome. Seamless integration of big data genomics, phenomics and climate analytics can accelerate precise identification of regions on crop genomes responsible for variations in phenotypic traits, and narrowing down to promising candidate genes for trait introgression by genome editing (Blake et al., 2016). Since 2013, a range of new features have been produced in crop plants (soybean, mustard, wheat, millet, corn, rice, tomatoes, etc.) through genome editing that would have been difficult and time consuming to achieve with conventional breeding or existing plant transformation methods. The regulatory aspects of genome edited crops are still uncertain, but broad support appears possible because the end product does not contain foreign DNA (Jones, 2015). Some genome edited crops have been exempted from GMO regulatory procedures in USA and Canada, so long as no pest sequences are included in the genome. China, Europe and other countries are likely to follow suit on a case by case basis (Sprinket al., 2016).

The state-of-art in big data driven plant breeding is the convergence of big data of climate, genomics and phenomics, on a single platform that interfaces with genome editing technologies. The scope to capture value through in silico modelling for speedier trait identification, and for accelerated breeding by genome editing has led to rapid proliferation of technology platforms by various startups (IP Pragmatics, 2016). Examples include Caribou BioSciences, Cibus, KeyGene, Precision Biosciences, Calyxt, etc. Agri-multinational companies (Bayer Crop Science, BASF, Dow Chemical, DuPont, Monsanto, and Syngenta)

are increasingly leveraging strategic alliances with the start-ups to license gene editing technologies and big data algorithms for novel germplasm development. The market for such platform based genome editing is growing at over 30% annually and is expected to reach US$ 315 million by 2020.The public sector has generally lagged the private sector in this innovation cycle. One significant public initiative is the collaborative GOBII (Genomic Open-source Breeding Informatics Initiative) project at Cornell University with focus on five staple crops – wheat, rice, maize, sorghum and chickpea (http://cbsuss05.tc.cornell.edu/gobii/).

(iii) Providing farm decision support for CSA through customized knowledge delivery: On farm, the main source of variation in crop productivity for a given variety is the weather. In addition, crop performance varies with spatially variable soils, inputs, and farming practices. In general, farmers rely on prior experience and expert advice for decisions on varieties to plant, planting dates, and managing field operations. Their sources of advice range from other farmers to public agricultural extension services, NGOs, agribusinesses, and input retailers. They pay significant attention to weather based agro-advisories, and some may seek periodic online interactive support from experts. However, much of the advice they receive is of generic nature, applicable to the broad macro environment than to their specific farms. Further, the advice is exclusively focussed on productivity and profitability. The two other key objectives of CSA, adaptation and GHG reduction are not included. A few recent studies have addressed all the three objectives of CSA but at relatively large spatial units of districts as the basic land units to prioritize among a suite of generic CSA technologies (Shirsath et al., 2017).

The practice of CSA requires farm specific knowledge on local climate conditions and risks and detailed knowledge of other conditions at farm level. This implies that each farm needs to be characterized with respect to climate uncertainties and risks, resources, and technologies and practices, to enable CSA decisions. The primary decision is the choice of climate resilient crop variety. Based on this, decisions subsequent decisions can be made on planting dates for climate adaptation, land management for increasing carbon sequestration, and technologies and practices for improving water and fertilizer use efficiencies to reduce methane and nitrous oxide emissions. The state-of-art of climate big data analytics allows crunching down climate uncertainties and risks in current and changed climate to the farm scale (section (i) above). Similarly, the state of art of big data genomics-phenomics permits breeding and identifying crop varieties resilient to climate risks expected on the farm (section (ii) above). Increasingly, remote sensing data at high spectral and ground resolutions (1-5 m), and other data sources like crowd sourcing and mobile sourcing, are becoming available to characterize and monitor farm soil and crop conditions regularly. These data, when used with crop models can generate farm/er-specific knowledge support for evaluating and recommending decisions on input use, soil and water management and crop management for CSA.

The state-of-art in farm-level characterization and decision tools for CSA is evolving rapidly along with innovations in computer power, software, remote sensing, mobile technologies, crop models, data analytics and technologies for site-specific management. These developments when used with insights from big data analytics on weather, soils and crop performance can support farmers make data-based operational decisions that will optimize yield, boost revenue and minimize costs and chances of crop failure (Rosenzweig et al., 2013; Capalboet al., 2017, Antleet al., 2017).

Monsanto and Climate Corporation have demonstrated the feasibility of field-specific customization of advice by integrating big data based climate predictions, genomics and phenomics for plant breeding, and farm production management models into a viable

business model (Bell et al. 2016). The model is built on a big data climate-soil-crop genomics-phenomics platform that provides field wise (one acre) priced, prescriptive advice on selection of crop hybrids, nitrogen management, and risk cover to farmers in over 15 million acres in USA. In another example, IBM, University of Georgia, NOAA and local agencies are experimenting with IBM's Deep Thunder super-computing technology to make more specific and accurate weather forecasts by 1 sq km grids for individual farmers in the drought-prone southwest Georgia, USA. Their approach breaks down atmospheric information and real-time weather forecasts into 10-minute chunks about 72 hours ahead, and determines precipitation amounts, intensity and soil infiltration for each grid cell at 10 min intervals. The resulting site-specific data is transmitted to farmers' desktops, laptops, tablets, or smartphones, which enables them to decide onfarm operations. The partnership also integrates soil-moisture sensor networks, variable-rate irrigation systems, and more precise weather data to better conserve water and other resources to lower water usage by as much as 15 percent. A third example is of Coca Cola's fresh orange juice. The product is based on use of big data of weather, crop yields, satellite images, regional consumer preferences, and detailed product quality data on 600 different flavors that make up an orange, and many other variables. Machine learning algorithms decide harvesting schedules and blending of juice from oranges from various sources, to maintain a consistent taste all year round, despite differences in seasons and practices on individual farms (Ransbotham, 2015). The developments in data technologies have also led to many spin-off technology startups for farm decision support (AgFunder, 2015) from public research institutions (NASA - Planet Labs, www.planet.com; Los Alamos National labs - Descartes Labs, www.descarteslabs.com) and technology companies (Google - Climate Corporation). The start-ups monitor individual farms with high resolution micro-satellite and other sources at high frequencies and leverage other high resolution farm data with big data tools to support on-farm decisions.

However, reducing GHG emissions from farm soils has received little attention, though it is a core objective of CSA. Agricultural soils contribute 37% of global agricultural GHG emissions, mainly as N_2O from nitrogen fertilizers and CH_4 from rice fields (Paustian etal 2016). N_2O fluxes are directly related to N input management and, on average, about 1% of the N applied to cropland is directly emitted as N_2O(basis for estimating N_2O emissions in IPCC GCMs). However, this value is too high for under-fertilized crops and too low for liberally fertilized crops (Shcherbak et al., 2014).Unlike carbon, N_2O has no terrestrial sink. Reducing emissions by managing soil microenvironment is the only mechanism for reducing N_2O emissions. Soil additives and fertilizer coatings that inhibit or slow nitrification and tillage and water management practices can reduce N_2O emissions (ICF International,2016). The potential for GHG mitigation by on-farm soil and water management is large, but the distribution and diverse nature of soils and production systems (rainfed, irrigated, etc.) present challenges in accurately assessing and reducing emissions on individual farms. Similar is the case with methane emissions from agricultural soils (largely rice fields). Key determinants of soil CH_4 fluxes include aeration, substrate availability, temperature and N inputs, all of which vary spatially and temporally.

Models of N cycling have been integrated into crop models to predict soil emissions and reduce uncertainties. The Agricultural Model Intercomparison and Improvement Project (AgMIP) has created an ensemble of global crop models and climate change assessments to evaluate strategies and provide farm level decision support to concurrently improve crop productivities, on, climate change adaptation and reduction of GHG emissions (Rosenzweig et al., 2013). Emissions data is also available from satellites (Hardwick and Graven, 2016), and ground measurement networks. The measured data when combined with high resolution climate, soils, remote sensing data and model outputs constitutes soils big data. Integrating

high resolution climate, genomics crop, soils and other farm data with crop models in a unified framework can lead to farm specific advisories that can meet all the three core objectives of CSA. A framework for such integration is presented in Fig 3.

4. Roadmap for leveraging big data for climate smart agriculture in India

In summary, a big data analytics based portfolio of scientific and analytical tools for strategic and tactical decision support on individual farms has evolved over the past decade to operationalize and scale CSA. Initially the evolution happened independently across the domains of climate change, genomics/phenomics, and farm decision support systems, with the GCMs providing the primary climate change signal information. Later, with increasing capacity for hyper resolution climate and farm information, the portfolio of data and tools began to be integrated into implementable frame works on single platforms to characterize and manage every farm from a CSA perspective. The framework allows: (i) characterizing each farm by expected climate change uncertainties and extremes, and resources; (ii) enabling choice of resilient crops and varieties for the farm; (ii) applying farm level decision support tools in real time to optimize productivities, incomes, climate resilience, and reduce GHG emissions.

The basic input to the agri-big data platforms are the coarse resolution, but validated, and freely shared public domain GCM/RCM data, climate data bases, natural resources data, and genomic databases. Without free access to this data, leveraging big data analytics to crunch information to precise data, decisions and outcomes for individual fields would not be possible. In India, public domain digital data sources include (i) IMD for coarse resolution (1^0 x 1^0; 05^0 x 0.5^0; 0.25^0 x 0.25^0) climate data, (ii) 250m resolution soil data from National Bureau of Soil Survey and Land Use Planning; and high resolution 1m to 250m land use data from National Remote Sensing Centre. In addition, there are several international public domain climate data sources from which data for India can be extracted. These include historical data and climate projections to 2100 from IPCC-CMIP5; soils data from FAO, and remote sensing satellite data spanning a range of spatial resolutions (1m to several 100s of Km) from a variety of sources on land use, vegetation conditions, soils (including soil moisture), and water resources.

On the other hand, the local data and big data analytics algorithms that enable data crunching to farm scales are generally proprietary, and are fast becoming increasing sources of new intellectual property (IP) in agriculture. This has motivated a number of new business models based on big data technology platforms that host massive national and global databases from multiple sources, and tools to derive knowledge products and insights that support farm level strategic and tactical decisions. These developments are rapidly making site specific digital agriculture the foundation for CSA.

As weather becomes increasingly volatile with climate change, the mobile or smart phone becomes the farmer's most crucial tool for CSA. Knowledge delivery through mobiles brings to small holder farming the same advantages as large scale mechanized precision farming on large farms. The public infrastructure to deliver precise farm specific knowledge advisories to every farmer in every village in India on a mobile is rapidly becoming available through the Digital India Network which scales to 250000 village Panchayats. This will provide the opportunity to map every field and create multi-dimensional village and farm specific databases through surveys, high resolution remote sensing, crowd sourcing and other means to characterize each field from the CSA perspective, before generating prescriptive advice. The network can effectively be a two way channel that provides not only prescriptive knowledge services to farmers but also enables flow of data from every farm to the big data

technology platforms. These developments have made India a fertile ground for big data driven digital agriculture.

Thus, progressing towards a vision of connected, data-driven, and customized (to farm and farmer) climate smart agriculture in India is possible with the present state-of-art data science, technology, and systems approaches. But it is contingent on creating a national big data innovation ecosystem that combines heterogeneous, dynamic, and distributed datasets and analytics tools that radically enhance knowledge discovery for implementation of CSA. But, the present public institutional landscape of the National Agricultural Research System in India will need to change to assimilate the massive data, implement the data technologies and tools, and deliver data driven field-specific agronomic knowledge on CSA to individual farmers on mobiles or smart phones. This will require newer institutional structures, systems, skills, and mindsets for agricultural technology generation and transfer. Some key issues to be addressed are:

i. CSA requires a paradigm shift in the technology generation process in NARS from an empirical research station based field experimentation to a data and *in silico* modeling driven approach that complements research station and farmer-field experimentation. To enable such a paradigm shift, the public systems of NARS need to prioritize: (a) creation of nationwide public cyber infrastructure providing access to climate, soils, crops and genomics databases from centralized databases or clusters to enable creation of localized knowledge products for CSA, and (b) increase the value of data through policies that promote its authentication, data sharing and management to design CSA practices..
ii. Recognize the complementarities of public and private big data for CSA. Public data from climate models downscaled to regional grids (25/50 km) are the basic input data for big data machine learning algorithms. The local farm and other circumstantial data and the machine learning algorithms developed to crunch out farm/field specific climate change assessments is usually private data. Similarly, the genomic data of major crops is available in public domain. But the corresponding phenomics data and the genomics big data algorithms that lead to gene sequence-trait associations for climate resilience are often in the private domain. The private data and the machine learning algorithms constitute a new category of rapidly emerging valuable IP in agriculture.
iii. The choice of GCM and downscaling methods from amongst a plethora of web accessible models available is central to effective practice and scalability of CSA. Addressing thisdilemma requires a concerted effort by the public systems to identify standard datasets from specified GCMs after a comprehensive and comparative evaluation of downscaled climate projections by both climate scientists and agronomists.
iv. The skill sets required to develop big data platforms and apply big data analytics tools are not normally available in NARS. These pertain to the domain of computer science and mathematics. Institutions of NARS need to develop capacities to interface and collaborate with specialists in these domains in universities and other national and state education and research institutions at least at two levels: (a) farm datastorage, validation and extraction for analysis by individual farms; (b) generate, validate, authenticate and convert analytics derived information to farm specific advisories in real time in a form that can be implemented by individual farmers.
v. The creation of algorithms that crunch down GCM scale information to localized scales and scaling of digital knowledge delivery to millions of individual farmers is more in the domain of private sector. The proliferation of startups in agriculture in recent years is indicative of the increasing private sector role in digital agriculture. Institutions of NARS need to develop capacities to seamlessly interface and engage with the private sector, particularly the emerging start-ups, to ensure that they in turn engage responsibly with the farmers.

vi. The NARS is the only source of authenticated knowledge for farmers. But for ensuring sustainable value creation for the farmer for CSA, the public systems' role and responsibility for validation of knowledge delivered to farmers from digital sources only increases in the digital agriculture era. Newer institutional frameworks for such authentication will be needed as traditional authentication systems based on field trials at multiple locations, publication after peer review, etc., are time consuming and difficult to scale.

To conclude, the vision of a connected, data-driven, customized (to farm/farmer), digital and climate smart agriculture is achievable with the current state-of-art data sciences, technologies and integrated agricultural systems approaches. But it is contingent on building policy and institutional environments in NARS that promote institutional and individual competencies for engagement across highly diverse biological, physical, chemical, mathematical, engineering and social sciences, and public and private institutions. Nonetheless, such an integration will be possible with judicious use of big data analysis platforms and Internet of things (IoT) protocols for addressing different agricultural production systems under changing climate.

References

AgFunder (2016) AgTech Funding Report 2015, 59pp (https://agfunder.com/research/agtech-investing-report-2016)

Alder JR and Hostetler SW (2015) Web based visualization of large climate data sets, *Environmental Modelling &Software* **68**, 175-180

Altpeter F, Springer NM, Bartley LE, Blechl AE, Brutnell TP, Citovsky V, Conrad LJ, Gelvin SB, Jackson DP, Kausch AP, Lemaux PG, Medford JI, Orozco-Cárdenas ML, Tricoli DM, Van Eck J,Voytas DF, Walbot V, Wang K, Zhang ZJ, Stewart Jr CN (2016) Advancing Crop transformation in the era of genome editing, *Plant Cell*, **28,** 1510-20

Anandhi A, Srinivas VV, and Nanjundiah RS (2008), Downscaling precipitation to river basin in India for IPCC SRES scenarios using support vector machine, *International Journal of Climatology*, **28**, 401-420.

Andrade-Sanchez,P, Gore MA, Heun JT, Thorp KR, Carmo-Silva AE, Andrew BD, French AN, Salvucci ME, and White JW (2014) Development and evaluation of a field-based high throughput phenotyping platform, *Functional Plant Biology*, **41**, 68–79

Arthur WB (2011) The second economy, Mckinsey Quarterly, October 2011, pp 1-9 (http://www.mckinsey.com/business-functions/strategy-and-corporate-finance/our-insights/the-second-economy).

Assunção MD, Calheiros RN, Bianchi S, Netto MAS and Buyya R (2015) Big Data computing and clouds: Trends and future directions, *Journal of Parallel and Distributed Computing*, **79-80**, 3-15.

Antle JM, Jones JW, and Rosenzweig CE (2017) Next generation agricultural system data, models and knowledge products: Introduction, *Agricultural Systems*, **155**: 179–185.

Azmak O, Bayer H, Caplin A, Chun M, Glimcher P, Koonin S and Patrinos A (2015) Using Big Data to Understand the Human Condition: The Kavli Human Project, *Big data*, **3**, 173-188.

Bell DE, Reinhardt F, Shelman M (2016) The Climate Corporation, Harvard Business School Press, 44pp.

Benestad R (2016) Downscaling climate information, Oxford Research Encyclopedia, Climate Science (climatescience.oxfordre.com), Oxford University Press USA, 2016, 37 pp, DOI: 10.1093 / acrefore/9780190228620.013.27

Blake VC, Birkett C, Matthews DE, Hane DL, Bradbury P and Jannic J (2016) The triticeae tool box: Combining phenotype and genotype data to advance small-grains breeding, *Plant Genome*, **9**(2), 10pp
(https://dl.sciencesocieties.org/publications/tpg/pdfs/9/2/ plantgenome2014.12.0099)

Bomgardner MM (2016) Transforming agriculture, again, *Chemical and Engineering News*, **94**(34). 32-38.

Butler D (2013) When Google got flu wrong, *Nature*, **494**, 155-56.

Campbell BM, Vermeulen SJ, Aggarwal PK, Corner-Dolloff C, Girvetz E, Loboguerrero AM, Ramirez-Villegas J, Rosenstock T, Sebastian L, Thornton P and Wollenberg E (2016) Reducing risks to food security from climate change, *Global Food Security*, **11**, 34-43.

Capalbo SM, Antle JM, Seavert C (2017), Next generation data systems and knowledge products to support agricultural producers and science-based policy decision making, *Agricultural Systems*, **155**, 191-99.

Carbonell IM (2016) The ethics of big data in big agriculture, *Internet Policy Review*, Vol **5**, Issue 1, pp 1-13.

Chaturvedi R.K., Joshi J., Jayaraman M., Bala G., and Ravindranath N.H (2012) Multi-model climate change projections for India under representative concentration pathways, *Current Science*,**103, No.7,** 791-802

Chen S, Wu C and Yu Y (2016) Analysis of Plant Breeding on Hadoop and Spark, *Advances in Agriculture*, Vol 2016, Article ID 7081491, 6 pages http://dx.doi.org/10.1155/2016/7081491

Davenport TH (2014) Big Data at Work: Dispelling the Myths, Uncovering the Opportunities, Harvard Business Review Press, Boston, Massachusetts, USA, 240 pp.

Dhar V (2015) The scope of machine learning and deep learning, *Big Data* Volume **3** Number 3, 127-129.

Dhar V (2013) Data science and prediction. *Commun ACM*. 2013; 56:64–73

Edwards JD, Baldo AM and Mueller LA (2016) RICEBASE: a breeding and genetics platform for rice, integrating individual molecular markers, pedigrees and whole-genome-based data, *Database (Oxford)* , pages 1-6, doi: 10.1093/database/baw107 .

Ekström M., Grose, M.R., and Whetton P.H (2015) An appraisal of downscaling methods used in climate change research, *WIREs Clim Change,* doi: 10.1002/wcc.339

Faghmous JH and Kumar V (2015) Climate change: the case for theory guided data science, *Big Data*, Vol **2** No. 3, 155-163.

Fahlgren N, Malia AG and Baxter I (2015) Lights, camera, action: high-throughput plant phenotyping is ready for a close-up, *Current Opinion in Plant Biology*, **24**, 93-99.

Fan J, Han F and Liu H (2014) Challenges of big data analysis, *National Science Review* (China) **1**: 293-314

FAO (2010) Climate-Smart Agriculture: Policies, Practices and Financing for Food Security, Adaptation and Mitigation, FAO, 49 pp.

Fischer EM and Knutti R (2015) Anthropogenic contribution to global occurrence of heavy-precipitation and high-temperature extremes, *Nature Climate Change*, **5**, 560–564.

Ford JD, Tilleard SE, Berrang-Ford L, Araosa M, Biesbroekb R, Lesnikowskia AC, MacDonald GK, Hsu A, Chen C, and Bizikov L (2016) Big data has big potential for applications to climate change adaptation, *Proceedings, National Academy of Sciences,***113** 10729–10732.

Gandomi A, Haider M (2015) Beyond the hype: Big data concepts, methods, and analytics, *International Journal of Information Management* Vol 35, pp 137–144

Gilpin, L. (2014). How big data is going to help feed 9 billion people by 2050. http://www.techrepublic.com/resource-library/downloads/how-big-data-is-changing-farming-pdf-download/post/?skipAutoLoad=1

Ginsberg J, Mohebbi MH, Rajan, Pate S, Brammer L, Smolinski MS, and Brilliant L (2009) Detecting influenza epidemics using search engine query data, *Nature* **457,** 1012-1014

Girvetz, E.H., E.P. Maurer, P. Duffy, A. Ruesch, B. Thrasher, C. Zganjar, 2013, Making Climate Data Relevant to Decision Making: The important details of Spatial and Temporal Downscaling, The World Bank, March 27, 2013 (https://scholarcommons.scu.edu/cgi/viewcontent.cgi?article=1012&context=ceng)

Glendenning CJ and Ficarelli PP (2012) The relevance of content in ICT initiatives in Indian agriculture, International Food Policy Research Institute (IFPRI), Discussion paper 01180, IFPRI, Washington DC, USA, 40 pp.

Global Harvest Initiative (2014) The 2014 Global Agricultural Productivity Report, 65 pp, (http://www.globalharvestinitiative.org/gap-report-gap-index/2014-gap-report/)

Goly A, Teegavarapu RSV, and Mondal A (2014) Development and evaluation of statistical downscaling models for monthly precipitation, *Earth Interactions*, **18**, 1-20.

Halevy A, Norwig P and Pereira F (2009) The unreasonable effectiveness of data, *IEEE Intelligent Systems*, March-April 2009, pp 8-12.

Hardwick S., and Graven H.(2016) Satellite observations to support monitoring of greenhouse gas emissions, Grantham Institute Briefing paper No 16, March 2016, Imperial College, London, UK, 16 pp

Hendler J (2015) Data integration for heterogeneous datasets, *Big Data*, Vol **2**, No. 4, 205-15.

ICF International (2016) Charting a Path to Carbon Neutral Agriculture: Mitigation Potential for Crop Based Strategies, ICF International, 1725 I Street, NW ,Washington, DC 20006, USA, 145 pp.

IP Pragmatics Ltd (2016) Gene Editing technology: market assessment and Intellectual property Landscape, , IP Pragmatics Ltd, 74 pp.

Jackson E (2016) The value of big data in agriculture: inputs, farming and processing, Editor Introduction, *International Food and Agribusiness Management Review*, Special Issue - Volume **19** Special Issue A, 5-6

Jagdish, HV (2015) Big Data and Science: Myths and Reality, *Big Data Research*, Vol **2**, 49-52

Janssen SJC, Porter CH, Moore AD, Athanasiadis JN, Foster I, Jones JW, Antle JM (2017) Towards a new generation of agricultural system data, models and knowledge products: Information and communication technology, *Agricultural Systems*, **155**, 269-288

Jin X , Benjamin WW, Cheng X, Wang Y (2015) Significance and challenges of big data research, *Big Data Research*, Vol**2**, pp 59-64.

Jones HD (2015) Future of breeding by genome editing is in the hands of regulators, *GM Crops & Food*, **6,** 223-232.

Jones JW, Antle JM, Basso B, Boote KJ, Conant RT, Foster I, Godfray CJ, Herrero M, Howitt RE, Janssen S, Keating BA, Munoz-Carpena R, Porter CH, Rosenzweig C, and Wheeler TR (2017) Toward a new generation of agricultural system data, models, and knowledge products: State of agricultural systems science, *Agricultural Systems*, **155**, 255-268

Knowledge@Wharton (2014) Sustainability in the Age of Big data, Special Report, 16 pp. http://knowledge.wharton.upenn.edu/special-report/sustainability-age-big-data/

Kole C, Muthamilarasan M, Henry R, Edwards D, Sharma R, Abberton M, Batley J, Bentley A, Blakeney M, Bryant J, Cai H, Cakir M, Cseke LJ, Cockram J, de Oliveira AC, De Pace C, Dempewolf H, Ellison S, Gepts P, Greenland A, Hall A, Hori K, Hughes S, Humphreys MW, Iorizzo M, Ismail AM, Marshall A, Mayes S, Nguyen HT, Ogbonnaya FC, Ortiz R, Paterson AH, Simon PW, Tohme J, Tuberosa R, Valliyodan B, Varshney RK, Wullschleger SD, Yano M and Prasad M (2015) Application of genomics-assisted breeding for generation of climate resilient crops: progress and prospects. *Frontiers in Plant Science*, **6** Article 563.

Kune R, Konugurthi P, Agarwal, Rao CR and, Buyya R (2016) The anatomy of big data, *Computing, software practice and experience*, **46** 79-105

Magnin C (2016) How big data will revolutionize the global food chain, *Digital McKinsey* August 2016, 6 pp.

Maciejewski R and Douglas MC (2016) Visualization for data science: adding credibility, legitimacy, and saliency, *Big Data*, Vol **4**, No.2, 73-74

McKinsey Global institute (2013) Game Changers: Five opportunities for US Growth and Renewal, 172 pp

National Research Council. 2016. Attribution of Extreme Weather Events in the Context of Climate Change. Washington, DC: The National Academies Press, https://doi.org/10.17226/21852 .

National Science Foundation (2012) Core Techniques and Technologies for Advancing Big Data Science & Engineering (BIGDATA), Programme solicitation, NSF-12-499, National Science Foundation USA, 17 pp, https://www.nsf.gov/pubs/2012/nsf12499/nsf12499.pdf

National Science and Technology Council (2016); The Federal big data research and development strategic plan, Executive Office of the President, National Science and Technology Council, USA, 45 pp.,

Navarro P.J., Pérez F, Weiss J., and Egea-Cortines M.(2016) Machine Learning and Computer Vision System for Phenotype Data Acquisition and Analysis in Plants, *Sensors*, **16** 641; doi:10.3390/s16050641

Oertel C, Matschullat J, Zurba K, Zimmermann F, Erasmi S (2016) Greenhouse gas emissions from soils—A review, *Chemie der Erde* , Vol **76**, 327–352

Paustian K., Lehmann J., Ogle S., Reay D., Robertson G.P., and Smith P.(2017) Climate - smart soils, *Nature*, **532**, pp 49-57

Ransbotham S (2015) Coca-Cola's unique challenge: turning 250 datasets into one; MIT Sloan Management Review, 6pp, http://mitsmr.com/1SCkbiS.

Rehman MN and Esmailpour A (2016) A Hybrid Data Center Architecture for Big Data, *Big Data Research,* **3**, 29–4

Rose DC, Sutherland WJ, Parker C, Lobley M, Winter M, Morris C, Twining S, Ffoulkes C, Amano T, Dicks LV (2016) Decision support tools for agriculture: Towards effective design and delivery, *Agricultural Systems* **149**, 165–174.

Rosenstock TS, Lamanna C, Chesterman S, Bell P, Arslan A, Richards M, Rioux J, Akinleye AO, Champalle C, Cheng Z, Corner-Dolloff C, Dohn J, English W, Eyrich AS, Girvetz EH, Kerr A, Lizarazo M, Madalinska A, McFatridge S, Morris KS, Namoi N, Poultouchidou N, Ravina da Silva M,Rayess S, Ström H, Tully KL, Zhou W. (2016). The scientific basis of climate-smart agriculture: A systematic review protocol. CCAFS Working Paper no. 138. Copenhagen, Denmark: CGIAR Research Program on Climate Change, Agriculture and Food Security (CCAFS). Available online at: www.ccafs.cgiar.org

Rosenzweig C, Jones JW, Hatfield JL, Ruan AC, Boote KJ, Thorburn P, Antle JM, Nelson GC, Porter C, Janssen S, Asseng S, Basso B, Ewert F, Wallach D, Baigorria G, Winter JM (2013) The Agricultural Model Intercomparison and Improvement Project (AgMIP):Protocols and pilot studies, *Agricultural and Forest Meteorology* Vol **170**, 166–182

Rosenzweig C, Elliott J, Deryng D, Ruane AC, Müller C,Arneth A, Boote KJ, Folberth C, Glotter M, Khabarov N, Neumann K, Piontek F, Pugh T, Schmid E, Stehfest E, Yang H, and Jones JW (2014) Assessing agricultural risks of climate change in the 21st century in a global gridded crop model intercomparison, *Proc Natl Acad Sci, U S A.***111**3268-73

Runck BC, Kantar MB, Jordan NR, Anderson JA, Wyse DL, Eckberg JO, Barnes RJ, Lehman CL, DeHaan LR, Stupar RM, Sheaffer CC and Porter PM (2014) the reflective plant breeding paradigm : A robust system of germplasm development to support strategic diversification of agroecosystems, *Crop Science*, Vol **54**, 1939-1948.

Scheben A and Edwards D (2017) Genome editors take on crops: Genome editing technologies may help to enhance global food security, *Science*, **355**, 1122-23.

Schönberger VM and Cukier K (2013) "Big Data: A Revolution That Will Transform How We Live, Work, and Think", Houghton Mifflin, Harcourt Publishing, New York, 256 pp.

Shcherbak I, Millar N and Robertson GP (2014).Global meta-analysis of the nonlinear response of soil nitrous oxide (N2O) emissions to fertilizer nitrogen. *Proc. Natl Acad. Sci. USA* **111,** 9199–9204.

ShriffinRM (2016)Drawing causal inference from big data, *Proc. National Academy of Sciences* **113,** 7308–7309

Shirsath P.B., Aggarwal P.K., Thornton P., Dunnett A (2017), Prioritizing climate-smart agricultural land use options at a regional scale, *Agricultural Systems*, **5**1 174–183

Singh A, Ganapathysubramanian B, Singh AK, and Sarkar S (2016) Machine learning for high-throughput stress phenotyping in plants, *Trends in Plant Science*, Vol. **21**, No. 2, 110-23.

Snijders, C., Matzat, U., &Reips, U. D. (2012). Big data: Big gaps of knowledge in the field of Internet science. *International Journal of Internet Science* **7** 1–5.

Sonka S (2016) Big data Characteristics, Big Data Volume 19, Special Issue A, International *Food and Agribusiness Management Review*, 7-13 .

Sprink T., Eriksson D., Schiemann J., Hartung F.(2016) Regulatory hurdles for genome editing: process- vs. product-based approaches in different regulatory contexts, *Plant Cell Reports* (2016) **35**:1493–1506

Steenwerth KR, Hodson AK, Bloom AJ, Carter MR, Cattaneo A, Chartres CJ, Hatfield JL, Henry K, Hopmans JW, Horwath WR, Jenkins BM, Kebreab E, Leemans R, Lipper L, Lubell MN, Msangi S, Prabhu R, Reynolds MP, Solis SS, Sischo WM, Springborn M, Tittonell P, Wheeler SM, Vermeulen SJ, Wollenberg EK, Jarvis LS and Jackson LE (2014) Climate-

smart agriculture global research agenda: scientific basis for action. *Agriculture & Food Security*, 3:11, 39 pp; http://www.agricultureandfoodsecurity.com/content/3/1/11).

Stephens ZD, Lee SY, Faghri F, CampbellRH, Zhai C, Efron MJ, et al.. (2015) Big Data: Astronomical or Genomical?. *PloSBiol* **13** e1002195. doi:10.1371/journal.pbio.1002195

Thrasher B, Xiong J, Wang W, Melton F, Michaelis A, and Nemani R (2016) Downscaled climate projections suitable for resource management, *EOS, Transactions American Geophysical Union* **94** (37), 321-323.

Tripathi, S., V. Srinivas, and R. Nanjundiah (2006), Downscaling of precipitation for climate change scenarios: A support vector machine approach. *Journal of Hydrology*, **330**(3-4): p. 621-640.

UNFCC (2015) Report on the structured expert dialogue on the 2013–2015 review, FCCC/SB/2015/ INF.1 1 (http://unfccc.int/resource/docs/2015/sb/eng/inf01.pdf)

USDA (2015) USDA Roadmap for plant breeding, USDA, March 2015, 36 pp.

Vandal T, Kodra E, and Ganguly AR (2017a) Intercomparison of machine learning methods for statistical Downscaling: the case of daily and extreme precipitation. arXiv preprintarXiv:1702.04018.

Vandal T, Kodra E, and Ganguly S, Michaelis, Nemani R, and Ganguly AR (2017b) Deep learning: generating high resolution climate change projections through single image super-resolution, arXiv preprint arXiv:1702.03126.

Varshney RK (2016) Exciting journey of 10 years from genomes to fields and markets: Some success stories of genomics-assisted breeding in chickpea, pigeonpea and groundnut, *Plant Science* **242** pp 98–107.

Voytas DF, Gao C (2014) Precision Genome Engineering and Agriculture: Opportunities and Regulatory Challenges. *PLoSBiol* 12(6): e1001877. doi:10.1371/journal.pbio.1001877

Wang T, Hamann A, Spittlehouse D, Carroll C (2016) Locally Downscaled and Spatially Customizable Climate Data for Historical and Future Periods for North America. PLoS ONE 11(6):e0156720. doi:10.1371/journal.pone.0156720

World Meteorological Organization (2013) The Global Climate 2001-2010: A Decade of Climate Extremes, WMO-No.1103, 118 pp

World Bank (2013).Turn Down the Heat: Climate Extremes, Regional Impacts, and the Case for Resilience. A Report for the World Bank by the Potsdam Institute for Climate Impact Research and Climate Analytics. Washington, DC: 254pp.

World Bank (2016) World Bank Big Data Innovation Challenge Join us in rethinking climate resilience through big data solutions; Challenge Handbook; 18 pp. (http:// www.bigdatainnovationchallenge.org)

World Economic Forum (2016) The Global Risks Report, 2016; 11th Edition, World Economic Forum, Geneva, Switzerland, 103 pp.

Wolfert, S. Ge L., Verdouw C., Bogaardt M-J.(2017) Big Data in Smart Farming – A review, *Agricultural Systems*, vol **153**, pp 69-80.

Xiong J, DingJ and Li Y (2015) Genome-editing technologies and their potential application in horticultural crop breeding, *Horticulture Research* 2, 15019; doi:10.1038/hortres.2015.19.

Zhang, J. Naik HS, Assefa T, Sarkar S, Chowda Reddy RV, Singh A, Ganapathysubramanian B, and Singh AK (2017)Computer vision and machine learning for robust phenotyping in genome-wide studies. *Scientific Reports*, **7**, 44048; doi: 10.1038/srep44048.

Figures

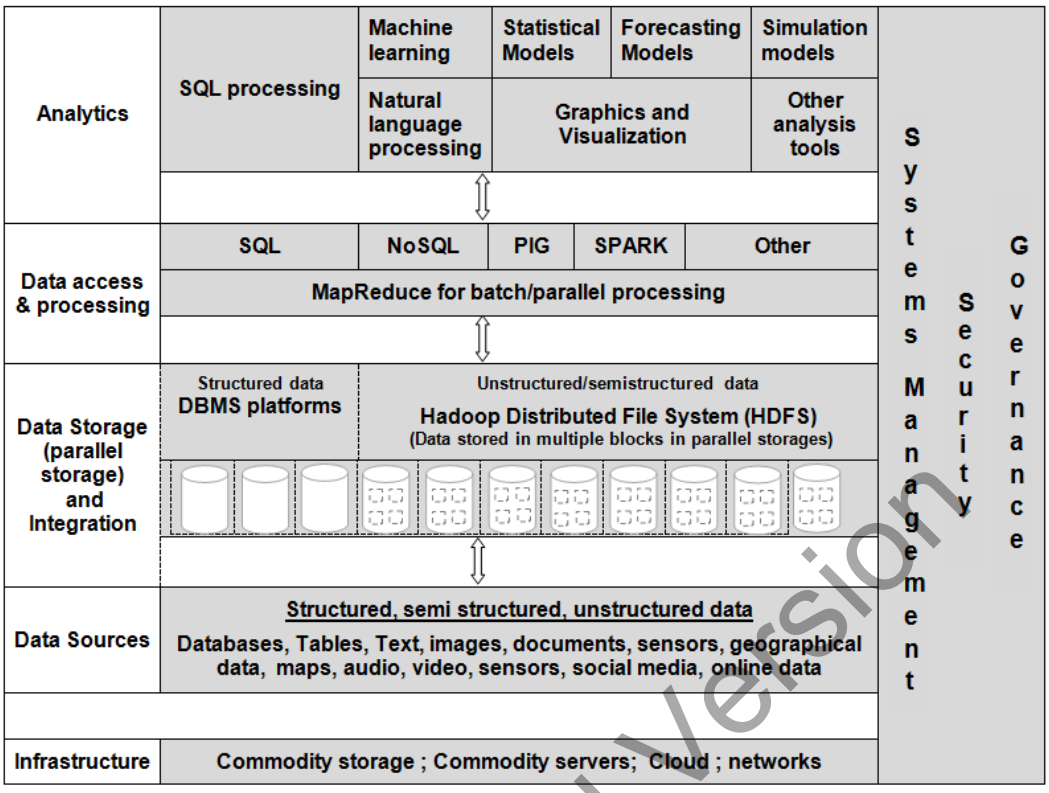

Fig 1: Big data ecosystem

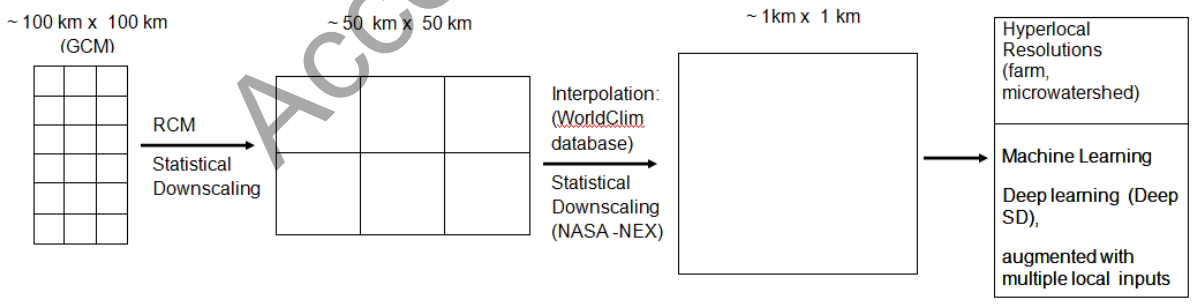

Fig 2: State of art of climate downscaling to hyper local resolutions

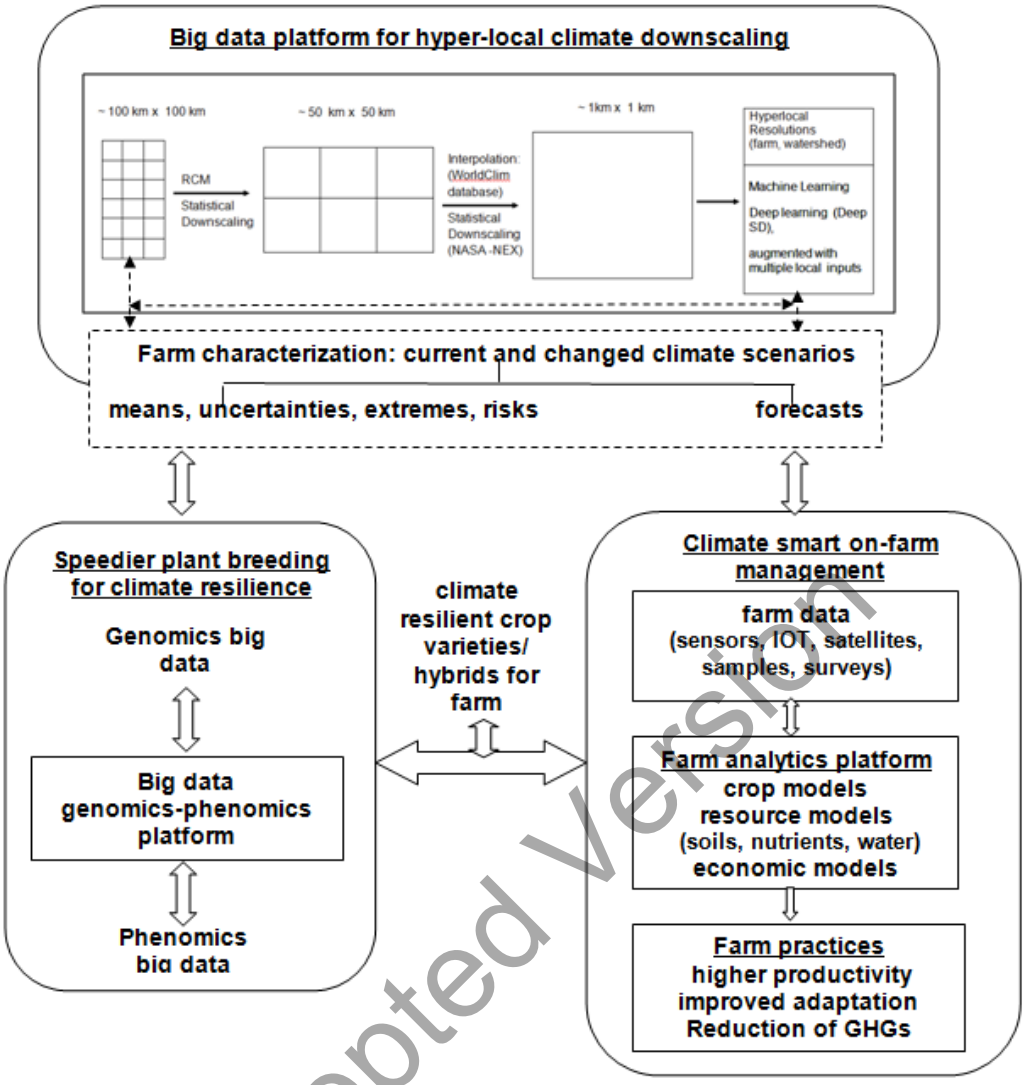

Fig 3: Proposed schematic architecture of the integrated big data analytics framework for climate smart agriculture