Weather-Driven Crop Statistics Enhancements for 21st Century Agriculture

A White Paper by Richard G. Stuff, Climate Assessment Technology, Inc.

Published March 21, 2013

Summary and Preface

Today's many and various early season estimates of crop production for the harvest ahead are immersed in large and undefined uncertainties about it's size. The uncertainties are a combined product of low accuracies in weather forecast data covering the remaining season and the models used to interpret that data – mainly into yields per acre.

The uncertainties become manifest to varying degrees throughout the season in terms of grain price volatility. Increased volatility in association with higher grain price levels over the last six years has become major food security concern because of its economic sway and potential to cause food shortages in almost any part of the world.

To curtail much of the speculative volatility associated with crop production uncertainty, this white paper advocates a transformative agricultural systems development project to attain and deliver superior resolution daily outputs at the county level. The accuracy goal for the yield estimates is to sufficiently qualify for use as official crop statistics. The system of models would integrate leading edge science and engineering technologies into a special resource center for mobilizing and bridging research, development, implementation and information dissemination.

An important reason for taking the crop statistics approach is the value and role of the USDA historical, county crop data. It's crisscrossing range of time and geography for proving models and model results is unique in the world, and adequate resolution of the infinite complex of factors "behind" its crop quantities justify a comprehensive and integrated building up of a working crop-system and supporting knowledge center.

Within the many research and development requirements – quantum improvements to crop systems models and seasonal weather forecasts are the principals. Each is a titanic challenge but thanks to accelerating advancements in computing technologies and tangential development efforts related to climate change and their associated national-global collaborative resource potentials, each is considered achievable in a two- to four-year intensive push-pull effort as coarsely described herein.

In addition to boosting basic information to help tranquilize market volatility, a set of better seasonal weather forecasts and interpretative models clearly would have far reaching benefits across all crop agriculture. Primary among them are crucial advancements of basic means for future agriculture's needs for improving sustainability and intensifying crop production – nationally and globally

Genesis of the Need and Challenge

Over the typical 100-day field life of a spring crop, it works to produce an average of 1 percent of its final yield per day – for US corn that's about 2 bushels per acre per day. A majority of days may be responsible for less than one percent and it may be totally indirect through vegetative growth. Some days may even take away from the 'net' that was 'made' over previous days, then a few near ideal growing days may account for 4 percent and compensate for "poor" days. Essentially, all the variability depends on the day's weather make-up and how that synchronizes and interacts with the hundreds of other factors in place as the complex dynamically unfolds across each "growing" point in each predefined area sample frame. Barring disastrous events that "wipe-out" acreage, the standing crop accumulates the daily yield-growth portions – those past and remaining ahead – into its final yield outcome – a quantity that can only be precisely measured after the crop is safely harvested.

From the food security standpoint, most stakeholders would like to know and adjust plans for the size of the crop beginning even before it is planted. As one harvest is finishing, its numbers and various estimates of how much the next crop will produce are continuously being inserted into seed orders, storage, trade, consumption and price projection formulas. Estimates of grain supply and withdraw sequences are in turn continuously reflected in market places through the mechanisms of futures exchanges, international trading, government policies, etc. Grain markets and marketing are made unique among commodities by the protracted period of weather induced supply uncertainties and other features that simultaneously "act" on prices within minutes of the related information being published. The U.S. corn and soybean crops will be used as context reference in this paper.

Prior to the beginning of each growing season, the USDA-ERS and many other organizations make 1- to 10-year production projections based on yield and acreage trends, macro economics, etc. The first significant seasonal revisions are usually in March when the USDA and others report results of prospective planting (acre) surveys. Unofficial seasonal revisions to yield estimates typically begin with planting progress reports – early plantings have more frost damage risk, rain delayed plantings often correspond to lower yields, etc. Further into the season as the crops become well established in more advanced vegetative phases, the acreage estimates become likewise "established" and officially represented by USDA reports at the end of June.

The most accurate in-season methods of estimating yields are based on objective field sampling and analysis of actual grain quantities. These methods are used by the USDA-NASS. They are labor intensive and only available after "grain" filling stages are reached – usually about a mid-season restriction. The first official USDA corn and soybean yield estimates are published in early August. Then as "grain" maturity progresses and data shifts to actual harvest (production) sources, revisions are published at subsequent monthly intervals until "semi-final" values are released in January of the next year.

Over the past 50 years a real conundrum has developed around the deficiency of information before and between the in-season USDA reports relative to its rapidly increasing value. The strains built by recent-year value trends are increasing due to price inelasticity associated with shrinking grain reserves (buffers), increased supply and demand variability, and their increasing need for just-in-time product scheduling throughout many of the "links" involved. The "problem" typically becomes most disconcerting between March and August when the grain prices, volatility, and option buy-sell volumes most often reach their annual maximums.

Also, many pre-harvest market information applications must now include world-wide supply-demand outlook dynamics. Lack of public information about the 1972 USSR wheat crop failure caused a major market disruption and losses for producers in other parts of the world. In response, the U.S. Government launched the "Large Area Crop Inventory Experiment (LACIE)" to help prevent such voids in the future by using satellite image and weather data to remotely monitor production (http://www.sciencemag.org/content/208/4445/670.abstract) . The project introduced several milestone, new technologies and concepts that have been transferred into crop monitoring systems currently in use around the world. And now, today's domestic in-season information uncertainties have some analogy to the international voids of 50 years ago.

Given relatively stable acreage estimates early in the season, the current domestic uncertainties originate almost entirely through yield uncertainties and their uncertainty " multipliers" that reside in the remaining-season (runout) weather. To fill yield information voids arising before and between USDA reports, dozens upon dozens of unofficial estimates are offered both publicly and privately. They are produced by methods that range from expert opinions, statistical relationships with many configurations of weather variables, a wide variety of crop "condition" indices, to highly elaborate system simulation model outputs. With equally varied quality and quantity deficiencies in data, knowledge, computing, and other basics, the current stock of "ad hoc" yield estimation methods by necessity contain large amounts of "improvising".

At current stages of the technology development, the pre-harvest estimates by the ad hoc methods might be considered almost as much a part of the problem as a solution. Mainly, they suffer from low accuracies when evaluated at local "yield forming" geographies – counties in the case of official US statistics. A rare few if any sources have "packaged" their estimates with statistically correct accuracy statistics. Most of the commercial estimation services advertise their numbers at aggregated levels wherein county errors mostly cancel out, but cases (years) can arise where they "cancel-out" less (unadvertised numbers) and relay "negative" value to client-users. The latter condition is especially true when source algorithms do not account for inherent biases in model estimates – the most common being underestimation of extremes and underestimation in general. The phenomena is particularly costly when actual country-wide yields are significantly above trend levels. It is also the likely motive for a Dec 5, 2011 WSJ article criticizing the USDA pre-harvest estimates for being too high (http://online.wsj.com/article/SB10001424052970203752604576641561657796544.html?mod=WSJ_article_forsub).

With the possible exception of JRC MARS reports, no current methods using "remote" technologies are known to be "hard-wired" into official early season crop yield statistics. Surveyed from a step further back, it is unlikely that any earlier-season, official crop yield estimates beyond the USDA August-equivalent quantities exist anywhere.

The high cost for "big" data, complex models, and massive computing requirements restrict the development of the most advanced "remote" technology methods to government entities or a "dozen" of the world's largest companies.

Then, there is nothing to prevent a company or competing government that develops a relatively more accurate estimation method from keeping the results private and profiting from the advantages. In some marketplace cases they could add to the conundrum by occasionally making their estimates public after internal use and time-values fade – another phenomena that can occur in less than a day.

At the end of 2012, market information systems and food security policies improvements still have not dampened volatile spikes in grain prices. Some are large enough to disrupt national and sector economies into shortages and new famishments. The persistent wide diversity of pre-harvest weather, yield and production unknowns and media "flashes" about them can explode into speculative contributions to price volatility at any time. The volatility aspect of post 2006 grain prices has raised the subject to top ranking of global policy concerns. The most recent G8 and G20 Agricultures Ministers place it in the top-four of listed "needs".

To outline weather and yield formulations for supra extending in-season production information, this paper contains a thumbnail draft of a project prospectus for building an estimation capability to reach an accuracy level that can distinctly qualify it for incorporation into official pre-harvest crop statistics. An example baseline target would be to at least obtain state level accuracies* in July that would be equal to the current USDA August values for at least 95% of state-year cases. This is estimated to be about two to four times the current average capability of the available yield models for these crops. The specific qualification criteria would have to be established by the USDA or other user organization.

The case can be made that there is now sufficient "base visibility" of information, data, supporting "environment" and justifying needs to undertake the required effort. The objective is to build an adjunct yield estimation system ($sysBld$)* that extends and reconciles with the existing USDA methods to provide for precursory and continuous daily updating of expected yields and acreage losses with each new day's report of actual weather and forecasts. The three main project components – applied research and development, validation and testing, data acquisition and processing – should use about equal amounts of resources. The supporting "environment vista" includes the current development trends in data, modeling, and computing capabilities.

* *provisionally defined as percents of the deviations of final-harvested yields from trend values that are correctly projected at the various pre-harvest dates.*

* *acronyms:*

$modelA$ *– local for adjunct crop-yield estimation model*

$sysBld$ *–, local for daily estimation system build (the project)*

other – (non-local) references should appear on first page of internet search listings

Imposed Solution

The core element needed for the $sysBld$ objective is a crop model ($modelA$) that can achieve the target yield accuracy in the most extreme cases where the runout weather forecast has little or no skill (crop projections would be based on latest status and "normal runout" weather). For the near term, most runout weather is likely to be derived from climate probability forecasts and they can be considered the second most success-critical part of the system. It is anticipated that yield uncertainties become about equally attributed to model inaccuracies and runout weather inaccuracies in early July for U.S. corn and soybean estimations.

If counted globally over all crops the number of individual yield models published in the last 40 years probably reaches well over a thousand. Based on the "final" variables that are assigned coefficients for yields, they can be grouped into four general approach types – blocked weather, crop or environmental index, simulated growth, or some combination thereof. Only a portion of the total – maybe a third – have been applied to large area yields. Only a few of the "third" have been subjected to statistically valid evaluation testing. From there, it is down to a rare couple of cases with statistically valid comparative testing of a particular model against one or more other models.

Beyond the statistical tests that indicate the accuracy weaknesses, general model evaluations provide information with respect to both practical and theoretical constraints and improvement limits. For example, estimates based solely on index variables of "viewed" crop features have practically no season "run-out" projection capability – crop plants themselves do not "know" that they may abruptly run out of water or stricken by a blight within a few days time. Also, under conditions of changing genetics, climate, and management practices, index-yield relationships can be expected to be less correlated. This same trend applies to block weather variables having continually less synchronization with particular crop stages. Neither of these approach types have a feasible potential to provide the desired daily yield impact (change) metrics.

At a fundamental application level, $modelA$ needs to sufficiently reckon the growth-yield responses that correspond to dominant genetics, local soils and management practices in each area frame. Yields are basically nonlinear across decimeter space scales and hourly time scales and cannot be adequately represented by “broad-brush” averages that cover more than about one kilometer and one day. For other elements of the objective, it needs to sufficiently estimate quantities that correspond to USDA in-season observations of crop condition, growth measures, spectral indices, and other intermediary quantities.

Clearly, a growth-yield simulation that is incorporated in an appropriate crop system framework is the only viable core-model approach for meeting the objectives. Models using this approach can be constructed to have the dynamic data and information capacities for the needed variables and scales.

The framework implicitly contains fully validated and integrated phenology (see white paper), constituent soil moisture, planting rate, pest and other component sub-models.

Thus, alongside the $modelA$ core, $sysBld$ minimally includes the relevant sets or subsets of:

- soil science products for estimating the nutrient and water supply dynamics in the predominant soils used for the particular crops in the sample frames.
- atmospheric and climate science products for filing season-to-date and estimation of runout the weather variables.
- economic decision products for estimating variety selection, planting densities, fertilization rates, irrigation rates, and acreage abandonment under various degrees of damage by late frost, wind, hail, pests, or just low yields (to project yields to a end-of-season harvested acre basis) .
- insect, disease, and weed science products for estimating pest impacts in the sample frame areas.
- remote sensing products related to specific crop features of the target crops and/or their environment.
- statistical products related to accuracy assessment, model testing, calibration, parameter estimation, trend estimation, etc.

Since the subject and goals dictate the need for the comprehensive-robust-distributed open agronomic ecosystem framework, by necessity $sysBld$ is highly collaborative. These features also allow its “inputs and outputs” to include the array of ancillary crop information for linking and integration with other key monitoring, sampling data for verifications and feedbacks. They also open many channels for utilizing existing crop research and model components that are “qualified”.

The main requirements that make the to overall job so large include the range of scales (time and space), the infinite complexity of the systems, and the breath of scientific and engineering disciplines involved. Information for adding or fixing a model part cannot be obtained directly from county data or rarely even from field research plot data, but volumes of such data can be highly qualified to validate and test models. To attain the necessary performance improvements, model refinements will require meta-analysis types of syntheses of controlled environment and laboratory research results for building system frames that are as holistically and globally comprehensive as knowledge and resources permit.

Project levels of effort are thus determined by the current development status of system components relative to their 2- to 3-year $sysBld$ potential. The gaps range from small missing parts to intermediate “repairs’, to large constructs needed for both the models and data. The deficiencies may be most visible in the weather and climate spheres, but they are just as present in the crop and other information areas. The optimal system design will need access to maximum possible amplitudes of scientific knowledge, data and ICT – combined with transcending perspectives and innovations. Across the combined disciplines, resource allocation is anticipated to be about equally divided between model research, data preparation and processing, and validation and testing.

Extensive validation evaluations and statistical tests will be conducted prior to selecting core-model options and at every step in subsequent development processes. As the simulations and system become more sophisticated, new validation procedures may be required to allow detection of model flaws and vulnerabilities related to nonlinear, interaction, buffering, occasional synergistic effects, etc. Also, the most relevant and advanced statistical procedures must be rigorously invoked for accuracy assessments, comparisons of alternative models etc. Resolute validation and testing functions are critical for assuring the necessary model improvements and knowledge utilization in $sysBld$ components.

As well as being “open”, it is also anticipated that $modelA$ research and development will be greatly facilitated through collaboration in newly initiated programs like AgMip, CropM, iEMSs, etc. and facilitated by knowledge hubs and share programs like Agrimod, CropForge, Agropedia, etc. Overall, the core model and systems improvement objectives are considered feasible and within the time-line “striking distance” of achievement

The second large area of $sysBld$ resource dedication will be elicitation of better seasonal climate forecasts. The work is likely to consist mostly of evaluation testing of various forecast model capabilities (private sector and global government agencies, statistical, dynamical, global circulation models, etc.) and combinations of methods. As such, much of the required work would parallel that of NOAA’s Climate Test Bed and, like the other areas, the have a ready channel for collaboration. Other tasks involve the winnowing out the best possible weather simulation method to interpret climate probability forecasts. In turn, $sysBld$ should produce the best possible feed-back data and information about the complex regional crop surfaces to the weather and climate forecast development services.

The atmospheric and climate science product task area involves giant data base programming and management. Possibly, up to 50 years of historical daily surface weather data and perhaps 300 or more “pseudo years” of daily forecast weather may be stored in high resolution (approximately 1-km) grid formats. Each case may have 4 to 8 variables. Weather data for climate benchmark stations may require different storage structures and made relational to detailed point crop data. Likewise, many data collaboration opportunities should be present. Some possibilities may be the assimilation of research-plot data from sources like DataOne, JECAM, NEON, Sustainablecorn, etc. then with database engineering in programs like GeoShare, ESMF , SemaGrow, GEOSS, and cyber infrastructure developments in general. Other – even larger terabyte volumes of public data and available information are growing rapidly for parallel contributions and interchanges with $sysBld$ results.

Just as simulated crop features are produced to correspond to UDSA “ground” survey based data, they would be produced to correspond to spectral image data. The a priori simulated “images” can then be incorporated into the verification type analyses to enhance both classification (crop areas) and vegetation feature extraction capabilities, etc. Various analyses involving the USDA Cropland Data Layer is also anticipated. There are also many remote sensing projects such as GEO-GLAM, CGIAR-CSI for possible collaborations.

USDA crop district and county historical statistics are the central basis for $sysBld$. As a database of crop information, it is exceptionally endowed for its length, breath and overall quality. The extraordinary scope of weather, soils, crop varieties, management practices and other factors can be translated into statistical power when it is used in $sysBld$ evaluation testing, validations and calibrations.

Capacity and efficiency metrics of computer power continue to advance exponentially and put this requirement for full systems simulation modeling and validation analyses early within the project timeline. Possibly, $sysBld$ could utilize capabilities from Hypercomputing, Advanced Supercomputing, High-Performance Computing, etc.

Overall, completing an intensive, main project for US Corn and soybeans will require 4 to 5 years with 8-10 full-time professional-technical principals. In this process $sysBld$ becomes a fully rounded and integrated development center utilizing:

- collaboration with the most advanced public and private sector institutes and technical community leaders.
- managed at a special agronomic-meteorology-systems engineering facility but possibly staffed from anywhere
- linkage to an base university that is main source of project scientific, technical and management resources
- multiple implementation options and “distant” resources – possibly including partnering with an organization officially representing a major foreign production region and/or private organizations

With these ingredients, the project is initially estimated to be in the 50 million dollar range. An ideal funding would have a primary grassroots stakeholder like corn and soybean producer organizations, then hopefully, the USDA and /or a foreign country participant. Project management would consist of employing advanced efficiency methods throughout.

Expected Outcomes

By 5 am every morning the $sysBld$ portal has an updated set of yield, area, and production estimates as well as estimates of crop stages, and visual crop condition ratings, soil moisture, harvest losses, the 24hr changes, or the change from any user selected previous date and statistical distributions for each variable. The main output quantities are presented in color coded interactive maps, supporting charts and downloadable tables aggregated to any geography of interest. An example portal-interface map can be viewed at the USDA's CropScape portal. Alternate dates for the outputs are selected through calendar "clicks".

In this circa 2018 scenario, weather will still be in crop news, but speculations about its impacts are mostly replaced with stories about rescheduling (when, where, how much, how sure) of crop inputs and outputs – from the distribution of seed stocks to shipping containers. There will be other market news to move day-to-day prices, but it can be anticipated that the impacts of trade announcements, floods, etc. will be rapidly and more precisely assessed in various economic models and the resulting price changes will be restricted to smaller, truer increments. Above all, the real-time grain information "playing field" is almost "level" such that small sellers and small buyers around the world can participate and benefit more fully from futures markets and other risk management devices.

In addition to the targeted remedial effects of a daily-weather "reality check" on excessive grain price volatility and general rewards of more and improved market information, the supplemental, secondary, and indirect benefits of the $sysBld$ are most likely to be many times more rewarding. The extras are a result of $sysBld$'s high resolution, internal detail (component models) and data that can be used to improve scores of standard applications and add some new ones. Much of the added value is attributed to the projects driving stepwise improvements in the models and in seasonal forecasts

The largest payoff from the better models and season forecast duo will likely originate through their transfer to field level intensification-sustainability applications. These benefits are made possible by the crop district (or "blocks" of crop districts in the near term) sized resolution (footprints) of seasonal rainfall forecasts and scalability of the system simulation models. By using the "whole" of a current season's climate dynamics in the appropriate set of models, temporal information can be merged with the spatial information (field maps) to add another dimension to precision farming. This contribution is magnified in light of the "new" genetics geared toward specific kinds of seasons. Forecast weather sequence effects can be incorporated into field management zone decisions to support valuable scheduling adjustments for seeding, fertilization, crop protection, irrigation and harvesting. Since the "precision" season dimension does not require GPS-VRA equipped machinery, it is more easily transferred and scaled around the world.

One of the special benefits of calibrating $modelA$ with the USDA historical data is the extensive differentiation of particular climate, technology, soil, and other factors constituting yield trends. Similar insights can be derived for trends in acreages, planting dates, planting rates, etc. When these results are incorporated into the system's economic models, geographical acreage shifts, yield-acreage interactions, etcetera can be more precisely quantified and future-projected. For example, the basic tendency for the largest increases in crop acreages to occur in counties with the highest yields as displayed in the historic data should forecast much better with a model that used physical and economic factors as compared to using just the "year" factor alone. On longer range-global scales, these types of improved resolutions in the models help reduce the chances for "mistakes" in climate change mitigation, adaptation, and abatement efforts in agriculture.

The additional benefits are as far reaching as imaginations can ask about crop production. The component crop-weather simulation models – whether new or improved – are versatile analytical "tools" for translating system data into more accurate and revealing information. Some of the major applications include risk management, actual and potential productivity assessments, efficiency analyses (water use, nitrogen use, etc.), ecosystem service computations, germplasm phenotyping, climate change impacts, environmental impacts, adaptation assessments, etc. These capabilities in turn can be "game changers" for general production enhancement program efforts like climate-smart agriculture, sustainable agriculture, yield gap assessment and grow more with less.

While the most direct paybacks come from the improved models, data and forecasts, there are several indirect and longer-term $sysBld$ benefits. As the project center grows to include more crops and production regions, it helps grow the global agronomic systems data, technology, and knowledge bases. The associated professional capacity and community building makes the center itself an invaluable national agriculture asset. The combined development process forms a basic agronomic knowledge “wiki” that for any given time and place should be as “true” as possible to what has been measured in nature or measured in crop experiments using nature.

Another invaluable project benefit is U.S. leadership assurance in both our agronomic and market information and knowledge. All the major grain producing countries have initiated projects to incorporate more technology into crop monitoring and forecasting. India’s FASAL project schema appears to have the most similarity to $sysBld$. The EU’s MARS system has been using agrometeorological models and satellite image technologies for several years already and recently announced GLOBCAST for extending capability outside the EU. China has also announced plans to develop global monitoring. If any reach the skill levels described here, it highly likely that it have used the USDA historical database.

Tallied as profits made and losses prevented in grain marketing decisions, it is conceivable that $sysBld$ returns could surpass investment costs in a couple years. Then, if the direct gains and returns are like the tip of an iceberg, the real wisdom of the large investment imposed by the nature of the yield estimation challenge is the shared “health” gained in food security. If corresponding $sysBld$daily outputs were incorporated into official USDA crop statistics, the interactive reconciliations and verifications in the combining processes would place the resultants beyond reproach and shut out most speculation arising from questions about production impact uncertainty or assessments related to a season’s weather.