Just right: predicting allele-environment interactions

Genotypes for Smallholders and Specific Niches
Improved Design of Mechanistic Physiological Experiments

Ann E. Stapleton
Department of Biology and Marine Biology
University of North Carolina Wilmington

Reproducibility?

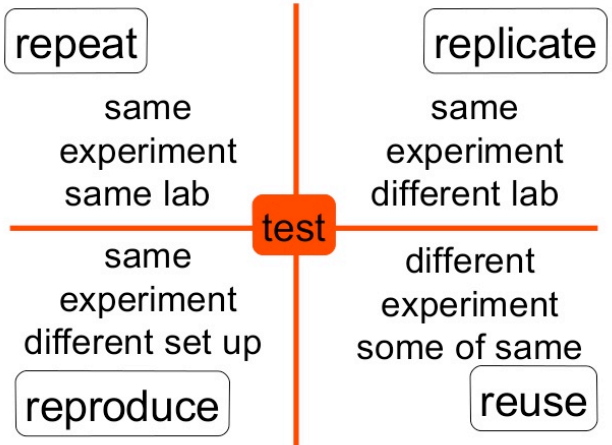
Review

Reproducibility in Science
Improving the Standard for Basic and Preclinical Research

C. Glenn Begley, John P.A. Ioannidis

Abstract: Medical and scientific advances are predicated on new knowledge that is robust and reliable and that serves as a solid foundation on which further advances can be built. In biomedical research, we are in the midst of a revolution with the generation of new data and scientific publications at a previously unprecedented rate. However, unfortunately, there is compelling evidence that the majority of these discoveries will not stand the test of time. To a large extent, this reproducibility crisis in basic and preclinical research may be as a result of failure to adhere to good scientific practice and the desperation to publish or perish. This is a multifaceted, multistakeholder problem. No single party is solely responsible, and no single solution will suffice. Here we review the reproducibility problems in basic and preclinical biomedical research, highlight some of the complexities, and discuss potential solutions that may help improve research quality and reproducibility. (*Circ Res*. 2015;116:116-126. DOI: 10.1161/CIRCRESAHA.114.303819.)

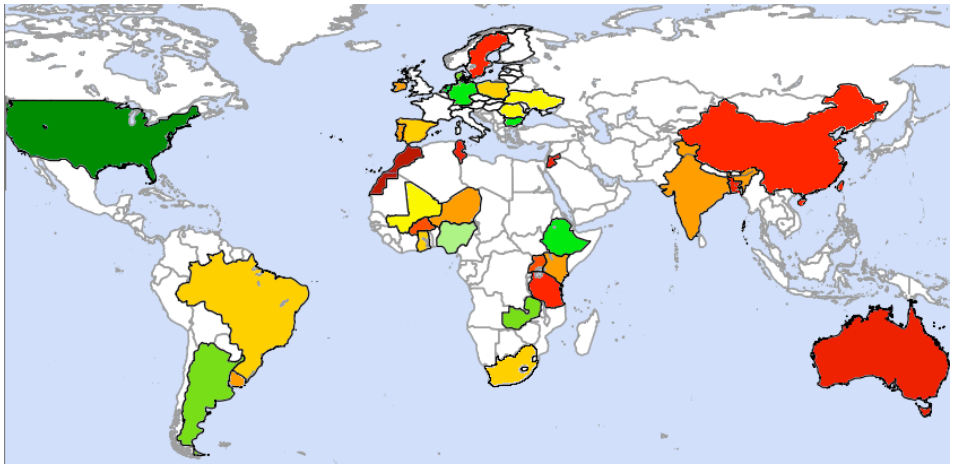
Key Words: funding ■ journals ■ research integrity ■ universities



Drummond C Replicability is not Reproducibility: Nor is it Good Science, online
Peng RD, Reproducible Research in Computational Science *Science 2 Dec 2011: 1226-1227.*

A. Shorten the cycle between explanation and prediction.

Yield gap

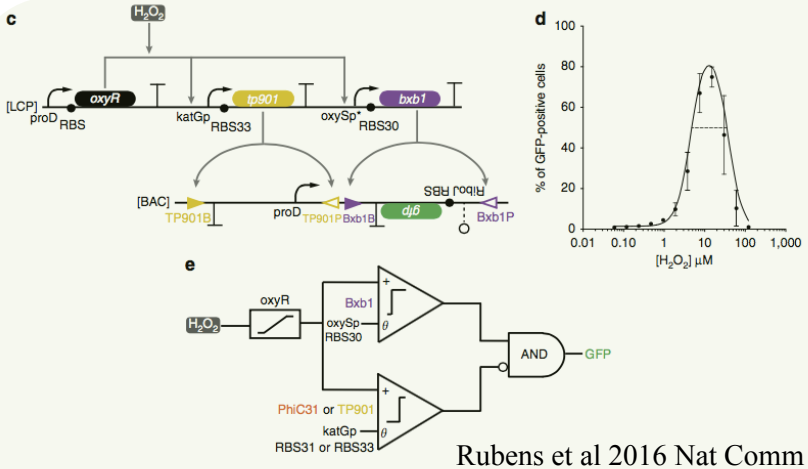


http://www.yieldgap.org/web/guest/yieldgaps

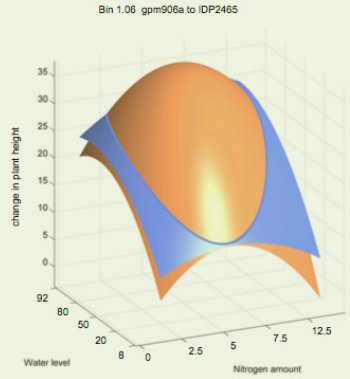
Why?

Regulation/homeostasis

Bandwidth filter
protein production, one signal

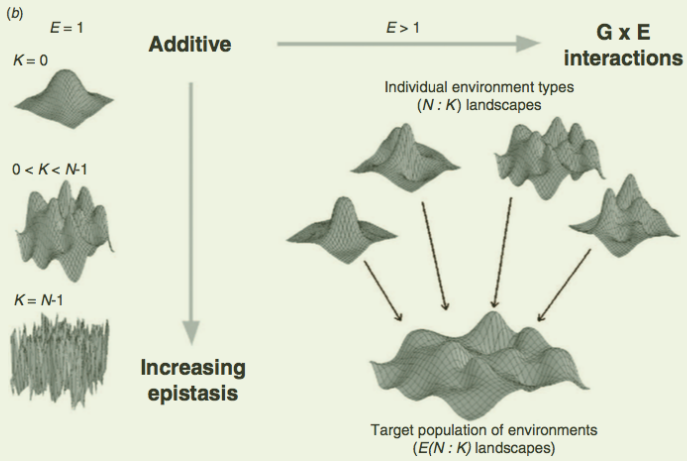


Rubens et al 2016 Nat Comm



The orange allele confers a more-than-additive response.

Genetic control of response surface
Alleles confer difference in growth as two inputs vary



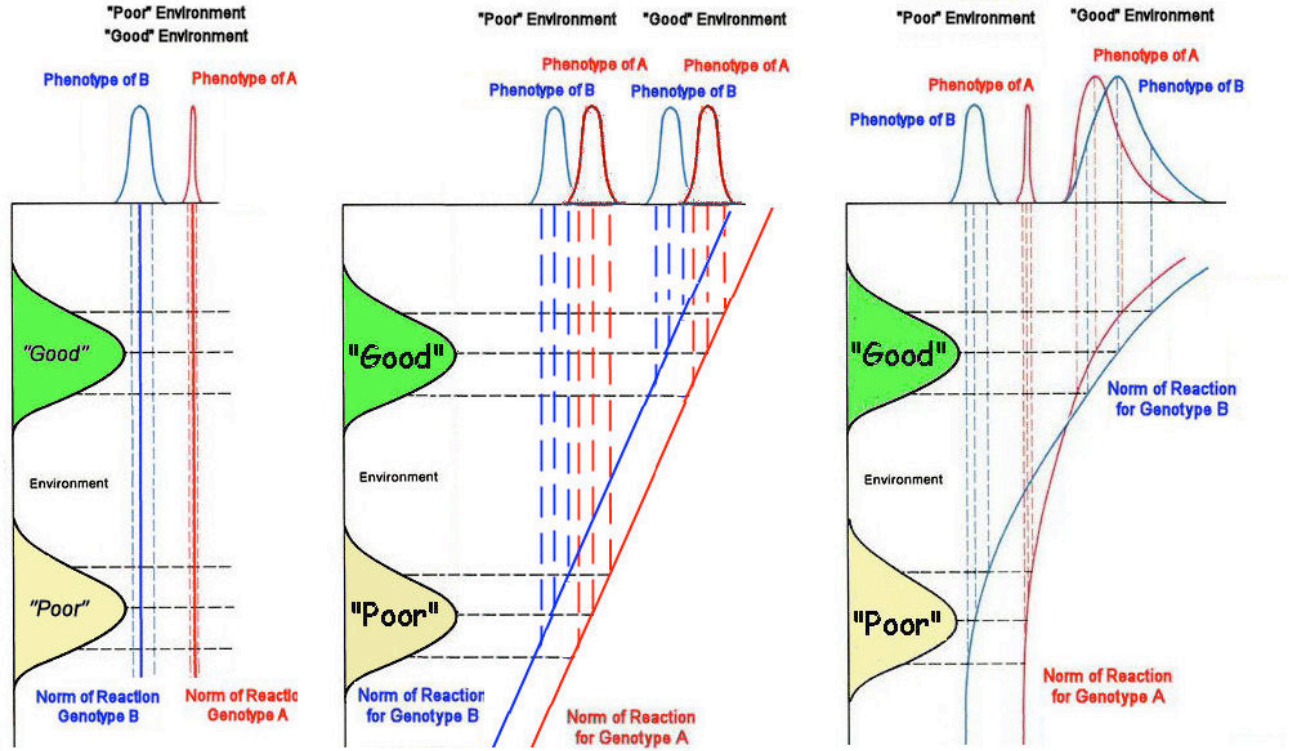
Fitness landscape for
G x G and G x E

Cooper et al 2005 Aust J Ag Res

Crossover genotype x environment

Additive, mostly genetic:

Mostly environmental?!
Environment hides genetics.

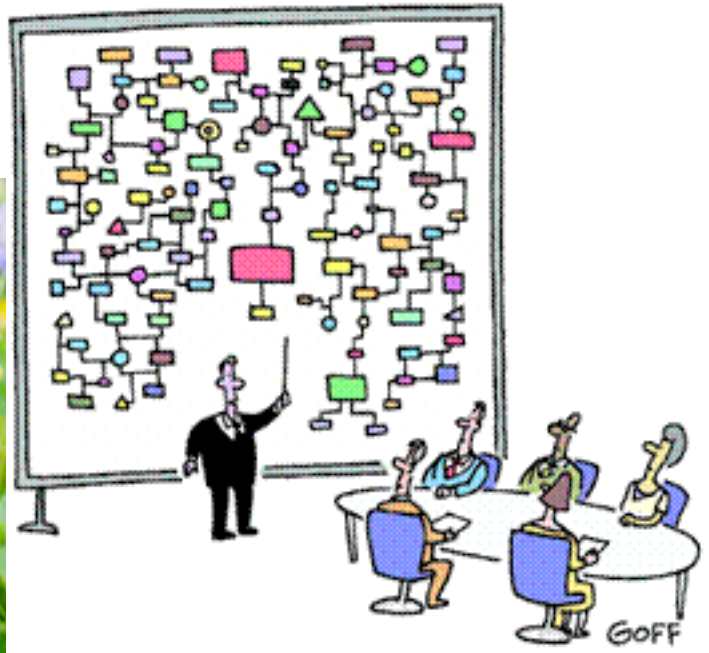


1997 Futuyma

Avoid or exploit?

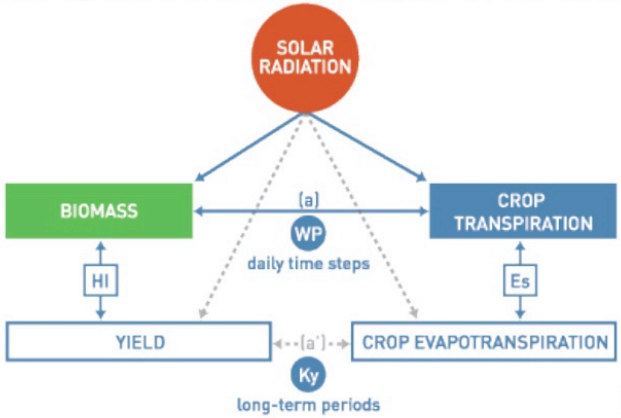
Large-scale breeding --artificial selection-- tries to reduce input variance and get large-area optima (parallel lines).

Multiple inputs

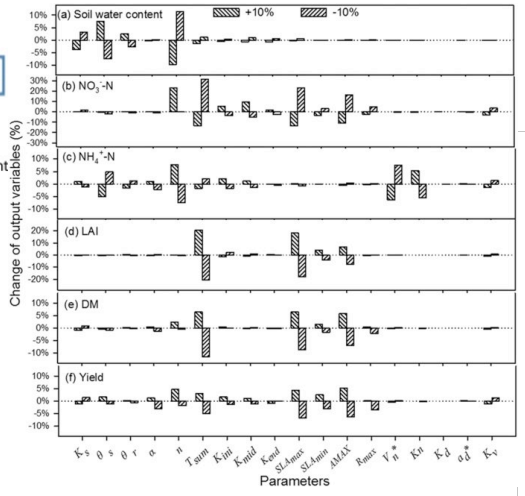


Crop Modeling – Bottom-up construction of crop production systems

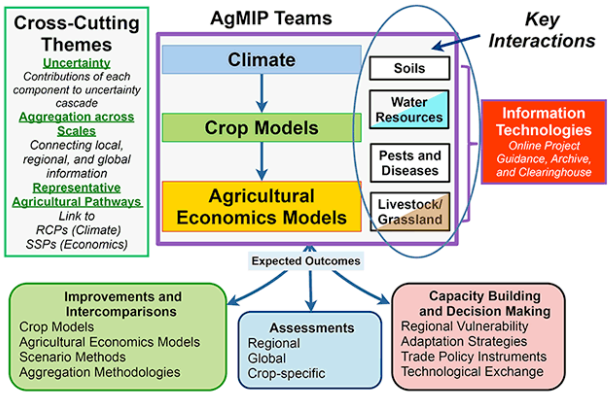
Components of AquaCrop, FAO model



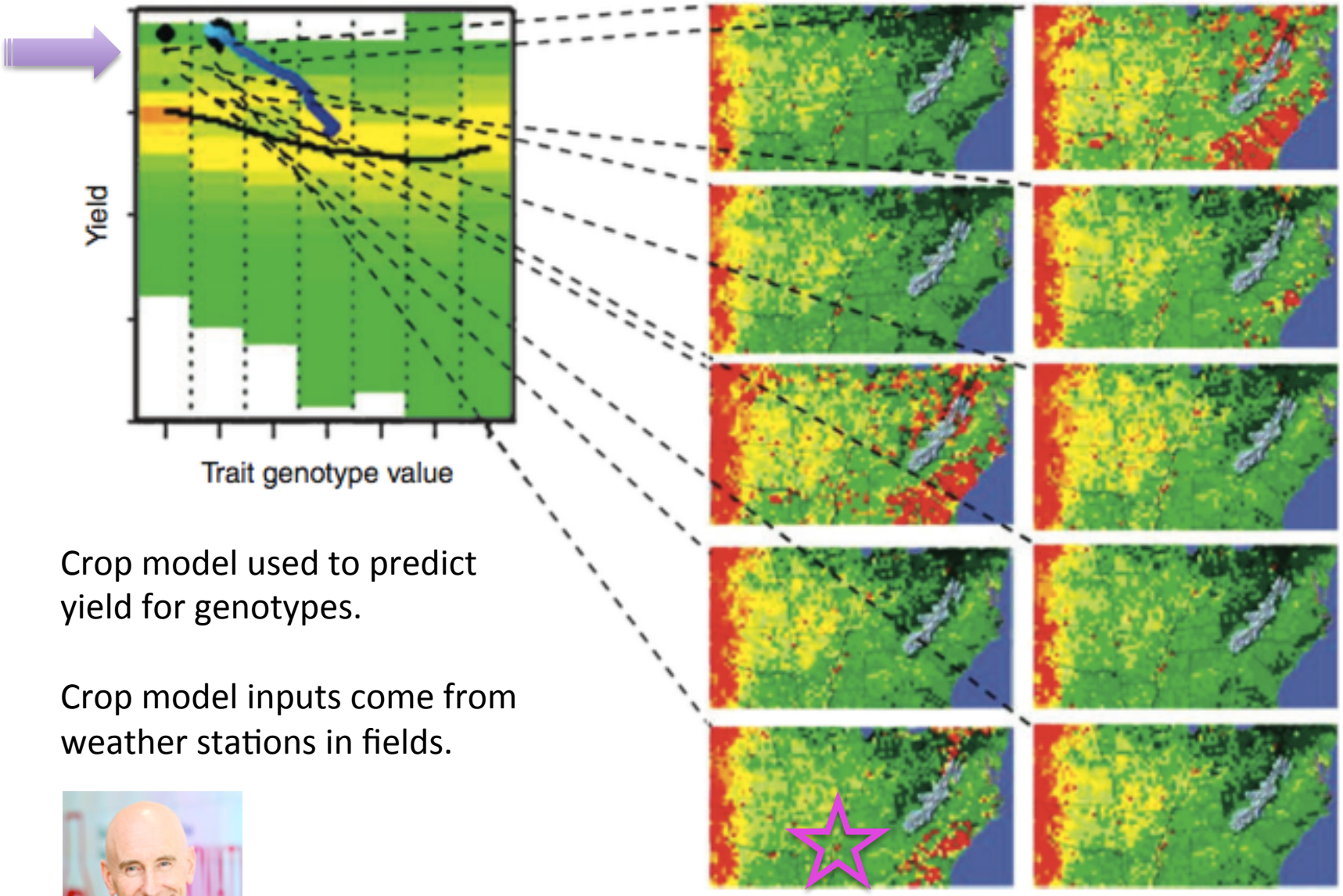
http://www.fao.org/nr/water/infores_databases_aquacrop.ht



AgMIP Teams, Linkages, and Outcomes



Rosenzweig et al., 2012 (in press)



Crop model used to predict yield for genotypes.

Crop model inputs come from weather stations in fields.

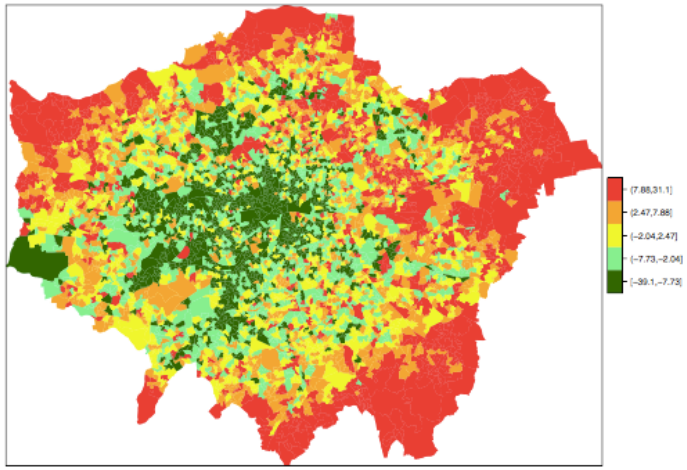
Cooper et al., "Predicting the Future of Plant Breeding" 2014 Crop Past Sci

Q. What makes the consistently red fields different from the other ones?

A. No single easily identified or easily modified feature.

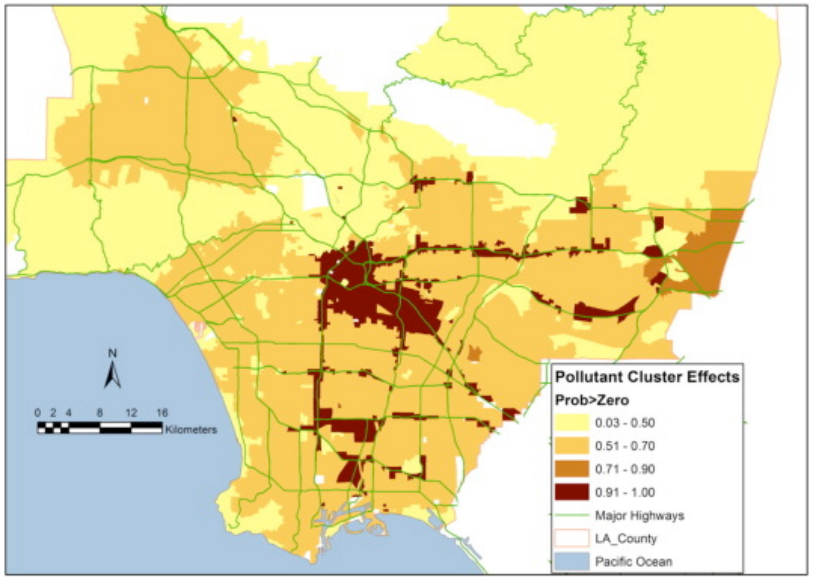
So how can we analyze this yield difference?

Spatial correlated response: deprivation in London



S. Liverani, Brunel University

Pollutant-Low Birth Weight, Los Angeles



Coker et al. Environment International, Volume 91, May 2016

Profile Regression

Dirichlet process Bayesian clustering, also known as profile regression

non-parametrically links a response vector to covariate data through cluster membership

Issues caused by

- correlated risk factors
- interacting risk factors

↓

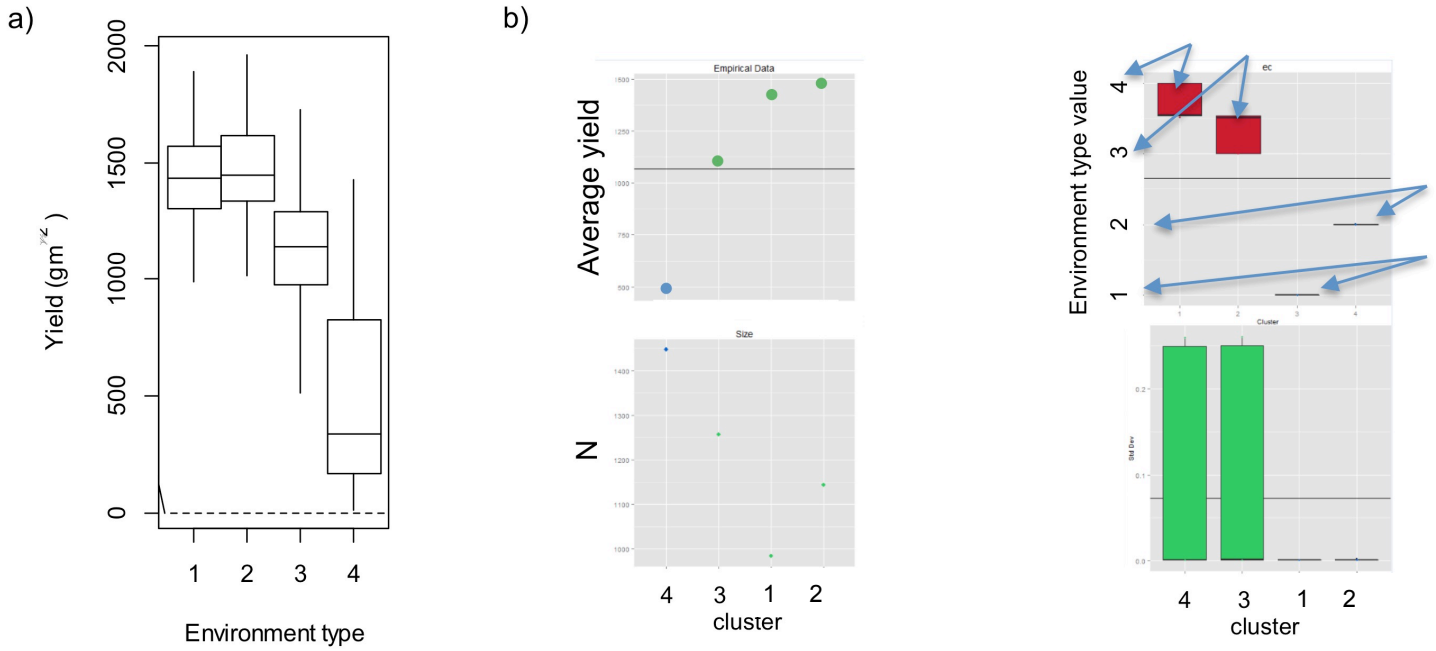
Profile regression

- **partitions the multi-dimensional risk surface into groups having similar risks**
- investigation of the joint effects of multiple risk factors
- jointly models the covariate patterns and health outcomes
- flexible but tractable Bayesian model

Profile regression effectively handles many correlated variables to generate useful predictors of an outcome.

Does it work for crop yield?

Tested using private Pioneer simulations:



Ongoing work: How much data is needed to get an effective prediction?

Profile Regression Example Result, Public Variety Trials

Experimental Design:

One year (2014)

Locations in Kansas, Missouri, Illinois (five fields per state)

Unbalanced design (not all genotypes/varieties in all fields)

Ten varieties analyzed

-- cannot use gene-level data from these commercial varieties

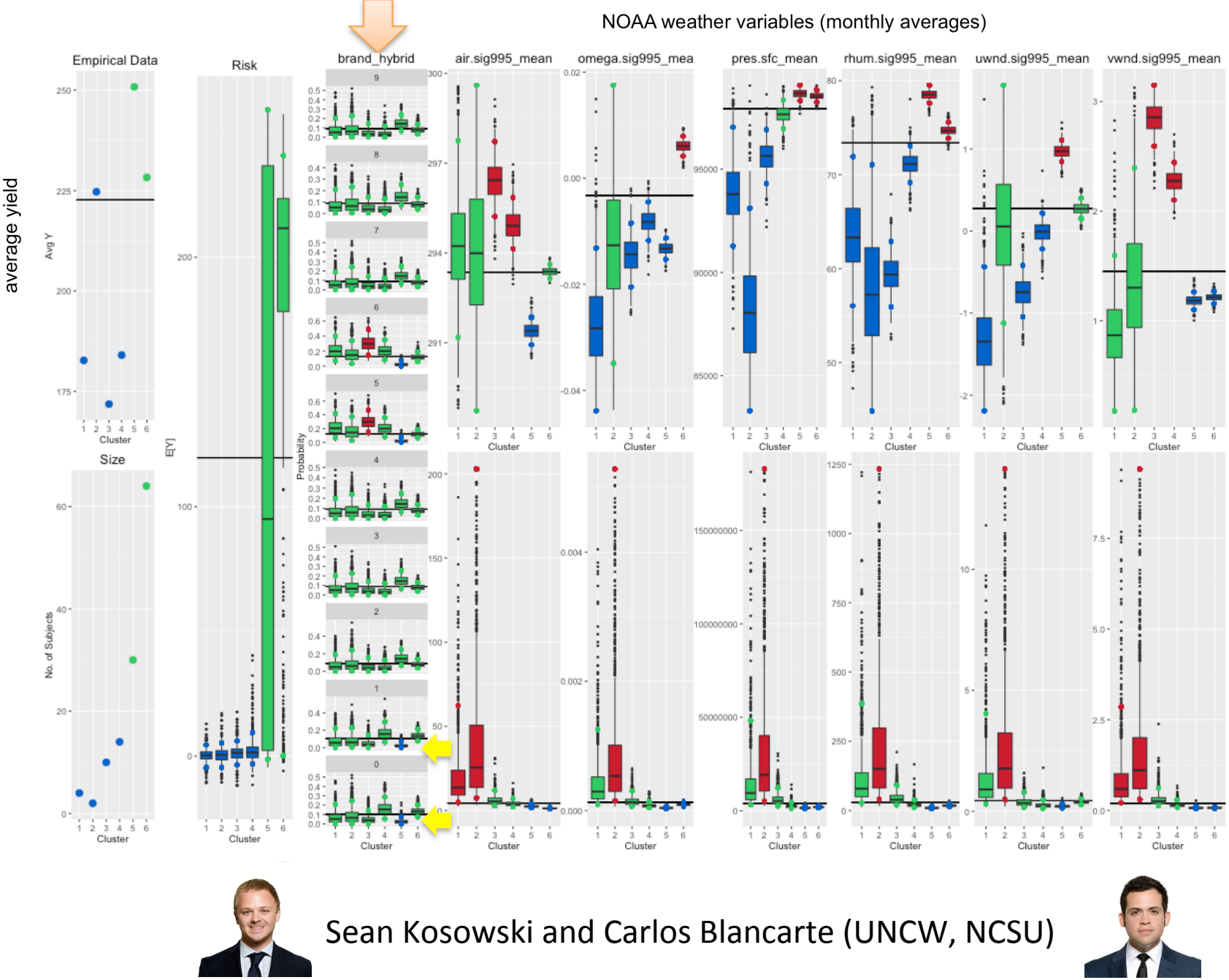
Weather covariates from NOAA

– monthly averages of six variables

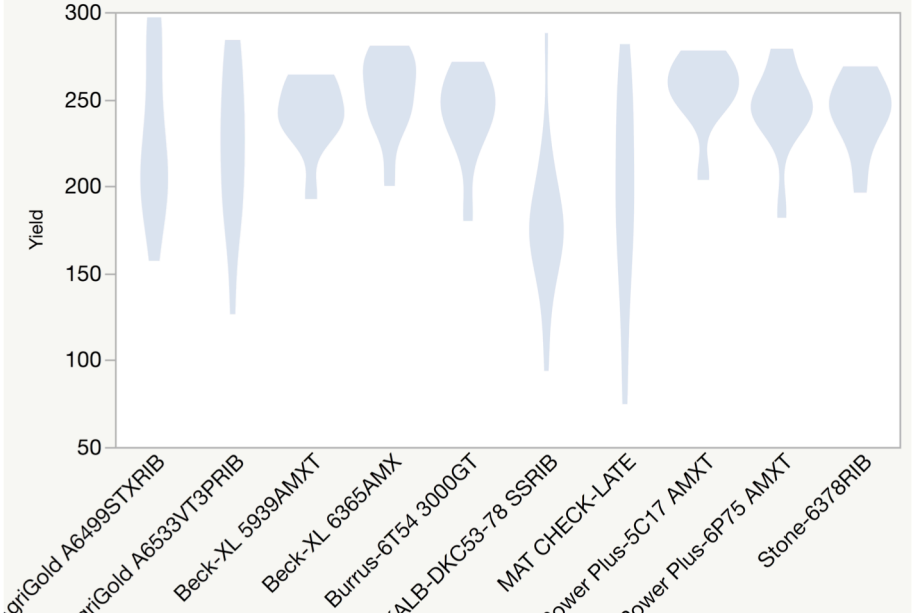
-- retrieved using approximate GPS coordinates, from nearest town

Which weather inputs matter?

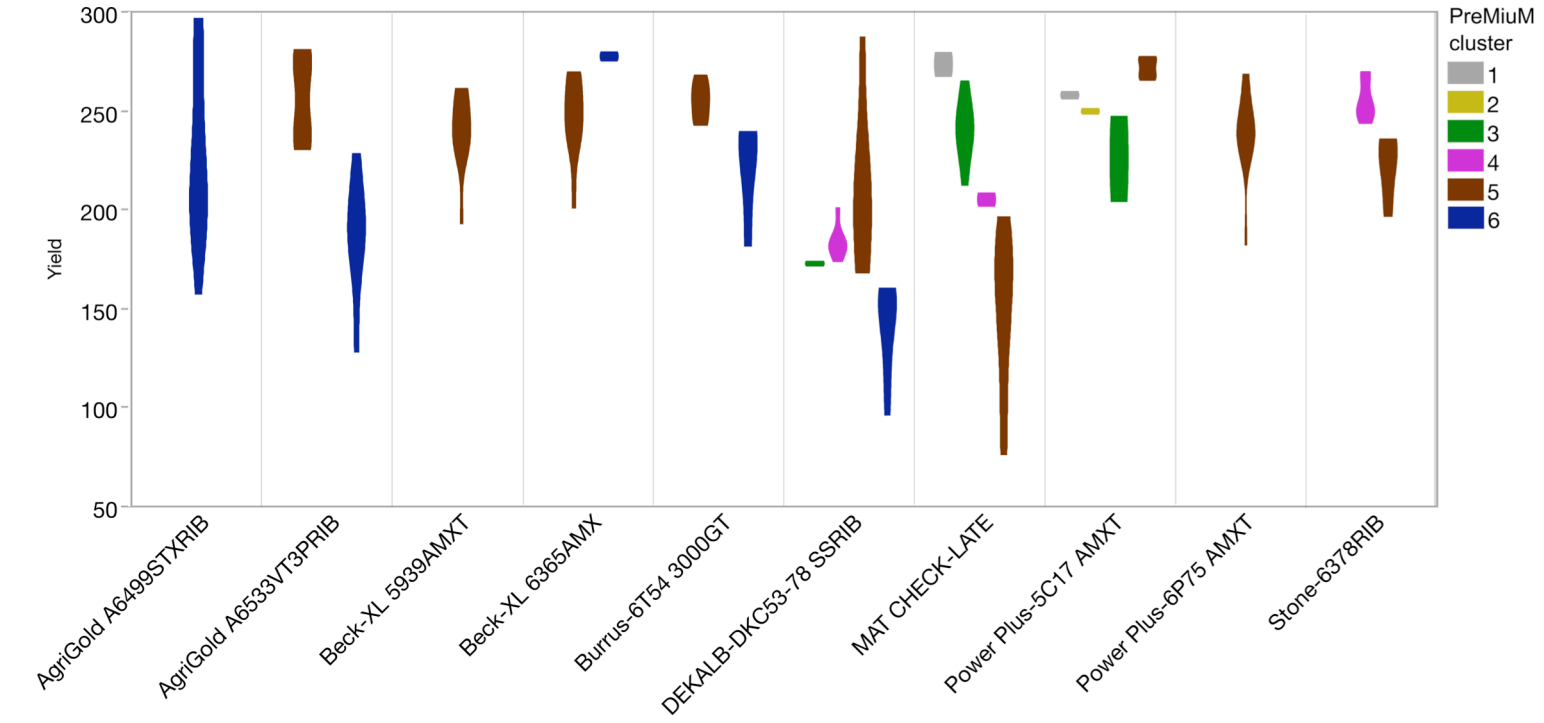
Do commercial genotypes vary across environments?



Sean Kosowski and Carlos Blancarte (UNCW, NCSU)



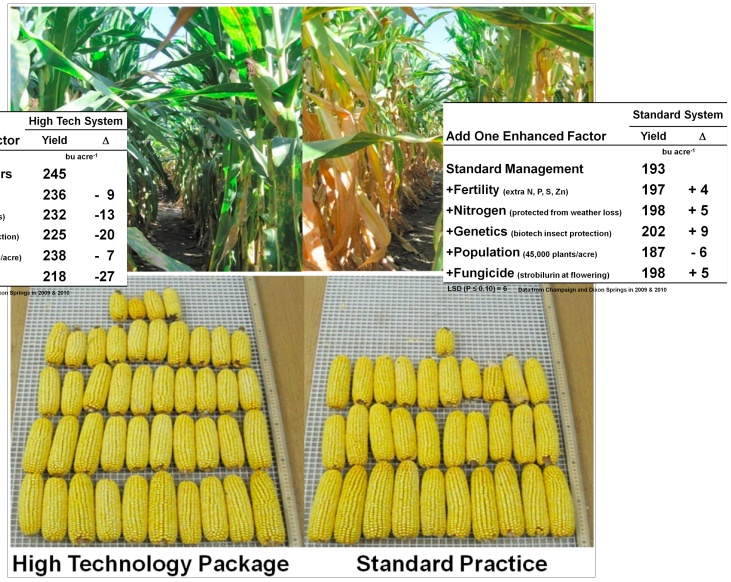
Which one would you choose?



Preview of Coming Attractions:

Cluster components could be used define experimental contrasts.

As a step in this direction, let's look at Below's omission plot analysis:



Omit One Enhanced Factor	High Tech System Yield (bu acre^{-1})	Δ
High Tech all Five Factors	245	
-Fertility (fertility from soil test)	236	- 9
-Nitrogen (unprotected from loss)	232	-13
-Genetics (-biotech insect protection)	225	-20
-Population (only 32,000 plants/acre)	238	- 7
-Fungicide (no fungicide)	218	-27

LSD (P ≤ 0.10) = 6 Data from Champaign and Dixon Springs in 2009 & 2010

Add One Enhanced Factor	Standard System Yield (bu acre^{-1})	Δ
Standard Management	193	
+Fertility (extra N, P, S, Zn)	197	+ 4
+Nitrogen (protected from weather loss)	198	+ 5
+Genetics (biotech insect protection)	202	+ 9
+Population (45,000 plants/acre)	187	- 6
+Fungicide (strobilurin at flowering)	198	+ 5

LSD (P ≤ 0.10) = 6 Data from Champaign and Dixon Springs in 2009 & 2010

High Technology Package Standard Practice

The sum of the factors under the standard practice was 17 bushels/acre, yet the difference between the standard practice and the high tech package was 52 bushels/acre.

Used contrasting 'backgrounds', removing one factor at a time.

Positive interactions exist among the factors when used in concert.

Agronomic factors – has not varied weather.

Ruffo et al., 2015 Agron J
http://cropphysiology.cropsci.illinois.edu/research/corn_omission_plots.html

Summary

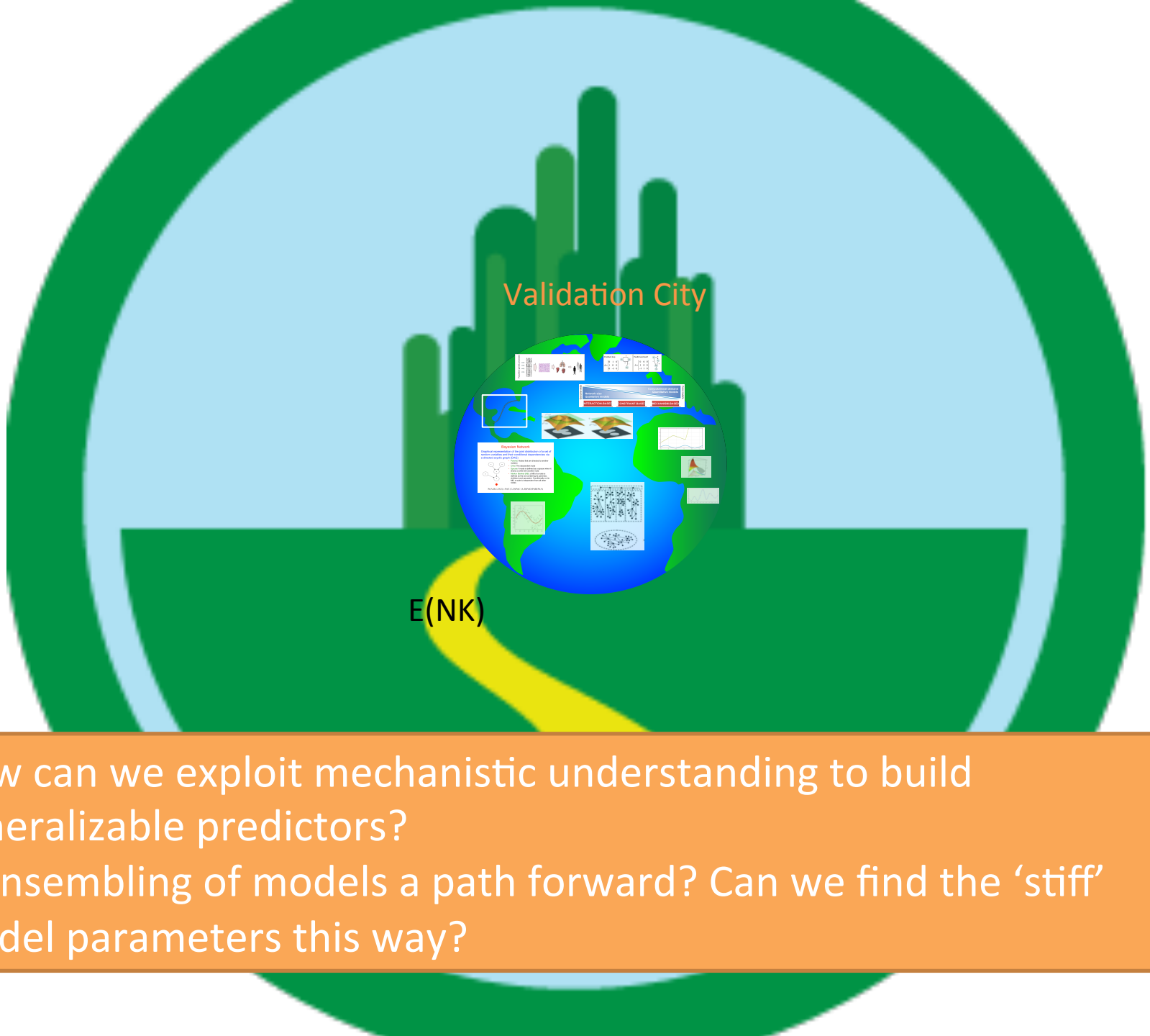
We can effectively group genotypes in weather covariate space using profile regression.

These ‘slices’ of weather provide new ways to select sites or design experiments.

Keep track of your covariates! Use your covariates!

We can design genotypes for niche markets and specific sites

-- for example, a smallholder’s field



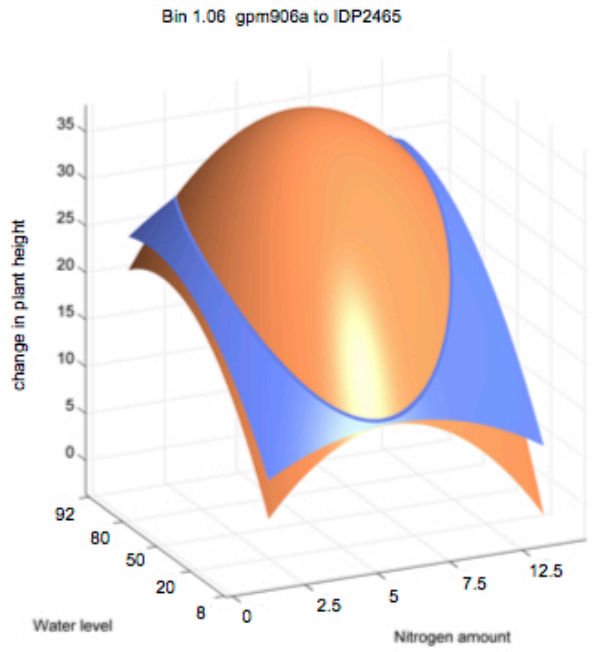
How can we exploit mechanistic understanding to build generalizable predictors?
Is ensembling of models a path forward? Can we find the ‘stiff’ model parameters this way?

Reproducibility implies mechanism:

"it should be physically possible to intervene on a putative cause variable in a mechanism without disrupting the functional relationships among the other variables in the mechanism"[1]

1http://plato.stanford.edu/entries/science-mechanisms/#WhaMecNotWhaNotMec

We identified loci with input combinations



The orange allele confers a more-than-additive response.

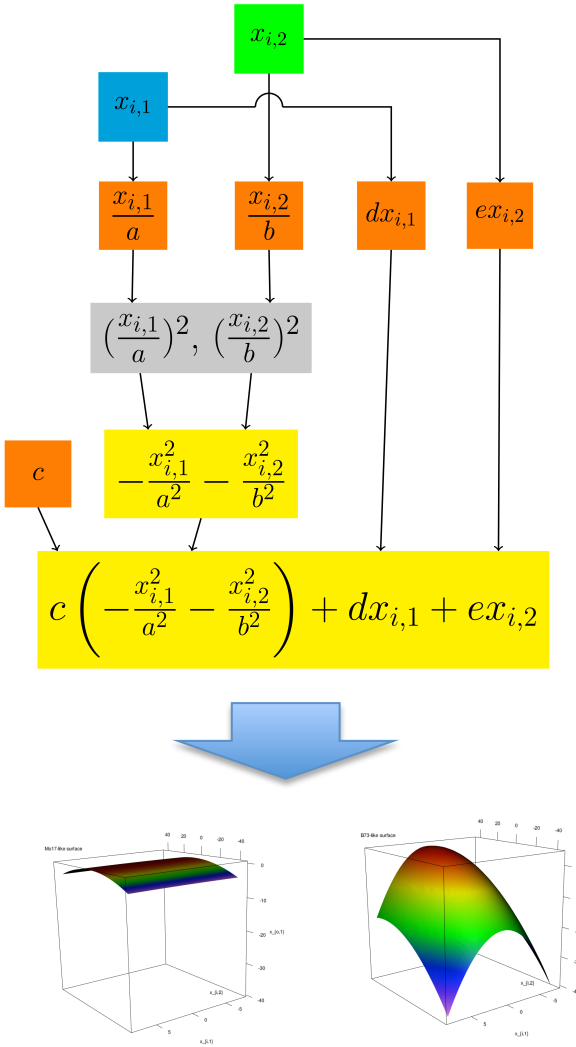
Will this keep happening? Would need interaction terms for each combination of each input!

Stapleton, Kazic, Chang, unpublished

We are currently fitting ‘producing functions’ by match to the response surface shape seen in the data.

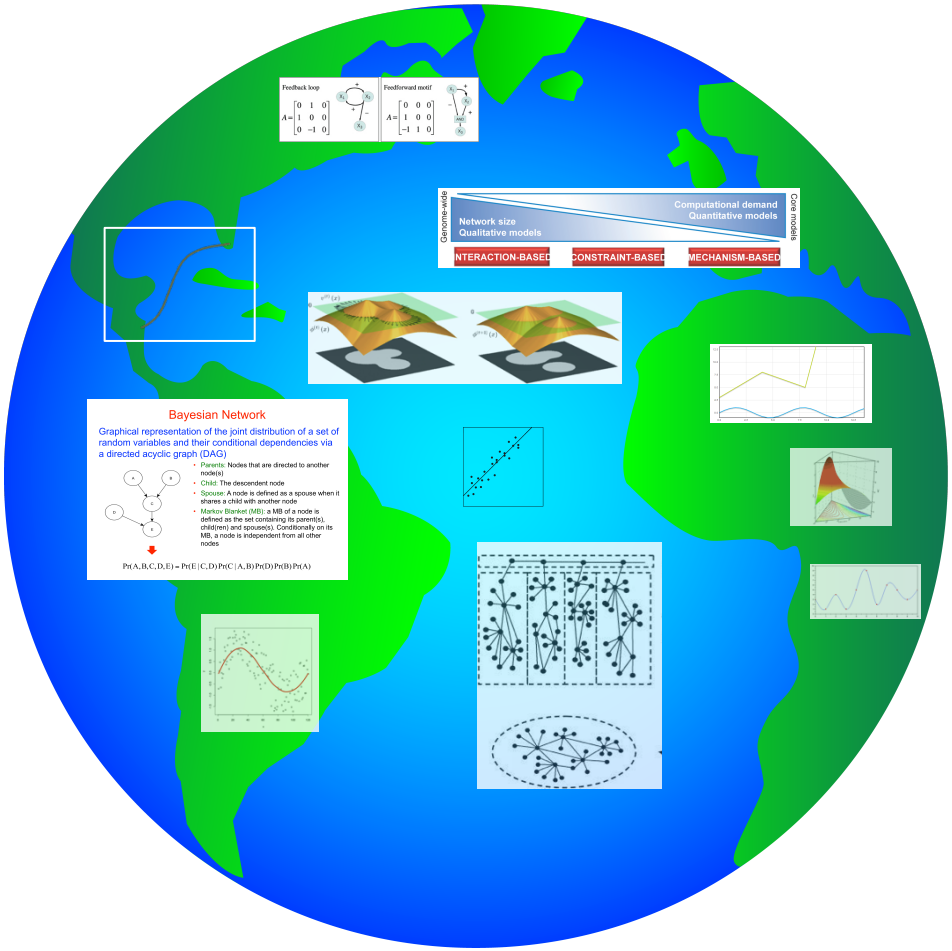
$$x_{o,1} = c\left(-\frac{x_{i,1}^2}{a^2} - \frac{x_{i,2}^2}{b^2}\right) + dx_{i,1} + ex_{i,2}$$

Note the negative, need this to bring the surface down after the peak

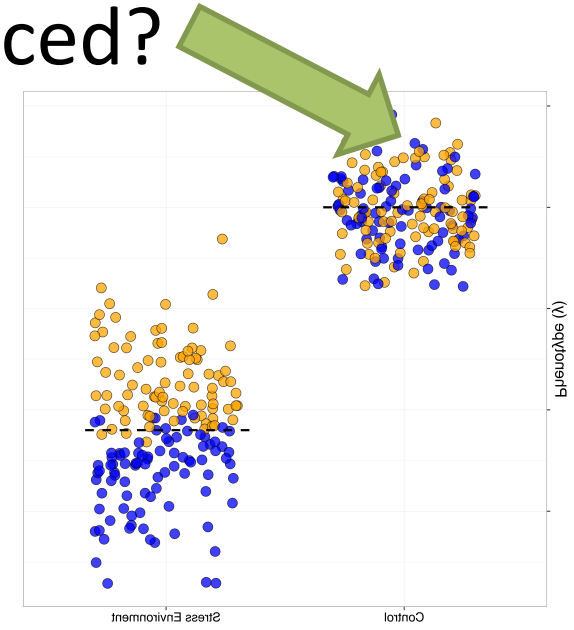


Stapleton, Kazic, Chang, unpublished
Inspiration from Babtie et al, 2014 and Eureqa

And then y = g + e + ge… and then



What about those cases when the spread of points is reduced, instead of enhanced?



knock-outs work, molecular biology is ‘unreasonably effective’

control inputs, benefit from selection for linearity/additivity and for tight range

For crops, this has lead to ‘hydroponic’ management to produce predictable g x g x m.

Spread of data points | Interactions: g, x e | Network behaviors | Network structures | Network i/o | Networks < -- > genes | Goldilocks and g x e | Producing functions

Point spread g x e | The power of linear equations | The potential in equation diversity