An Efficient Sensitivity Analysis Method for Spatio-Temporal Data from an Agricultural Systems Simulator

Bryan Stanfill David Clifford Henrike Mielenz Peter Thorburn

AASC 2014

Motivation

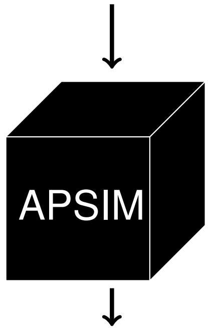
The **A***ricultural* **P***roduction* **S***ystems s***IM***ulator* is a widely used simulator for agricultural systems

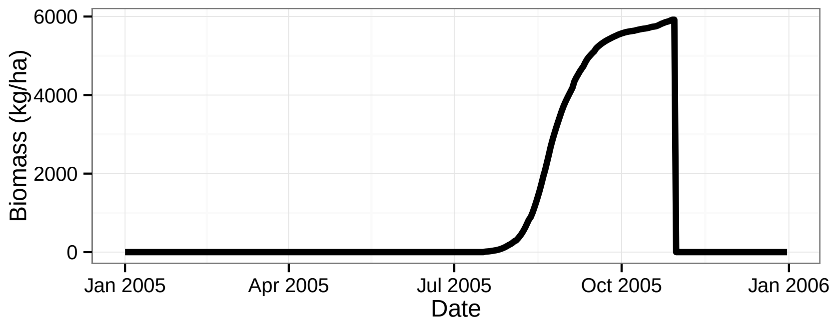
- Composed of several modules, each controlling a different aspect of the agricultural system
- Several sources of information are used: weather, farm management practices, soil and crop properties to name a few
- Deterministic - two simulations run with the same inputs will give identical results
- Dynamic - estimates are given over time
- More often then not the estimates produced by APSIM are taken at face value with little or no attention to uncertainty

Motivation

$$\text{grains\_per\_gram\_stem} = 25\text{g}^{-1},\ \ \text{max\_grain\_size} = 0.041\text{g},$$

$$\text{tt\_flowering} = 120^{\circ}\text{C days},\ \ \ldots$$

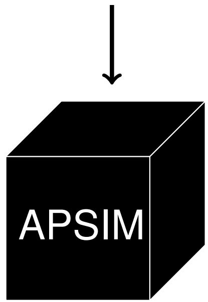




Motivation

$$\text{grains\_per\_gram\_stem} \sim N(25, 1),\ \ \text{max\_grain\_size} \sim N(0.04, 0.09),$$

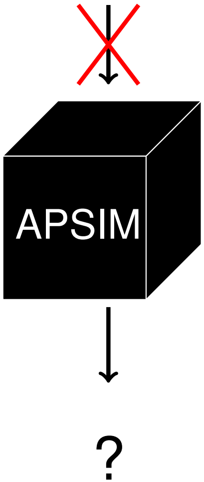
$$\text{tt\_flowering} \sim \text{Unif}(110, 130),\ \ \ldots$$



Motivation

$$\text{grains\_per\_gram\_stem} \sim N(25, 1), \ \ \text{max\_grain\_size} \sim N(0.04, 0.09),$$

$$\text{tt\_flowering} \sim \text{Unif}(110, 130), \ \ldots$$

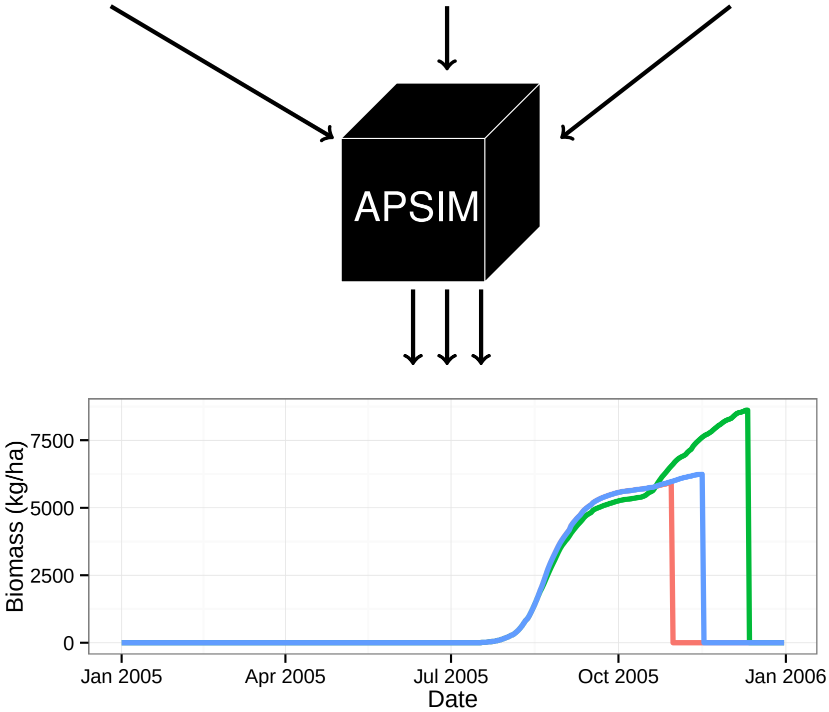


Sensitivity Analysis

- Identify how/which input parameters affect uncertainty in the output
- Provides insight to which parameters warrant further investigation
- The **first-order sensitivity index** S_i quantifies the proportion of variability in the output that can be attributed to its marginal relationship with input i
- The **total sensitivity index** ST_i quantifies the proportion of variability in the output that can be attributed to its complete relationship with input i

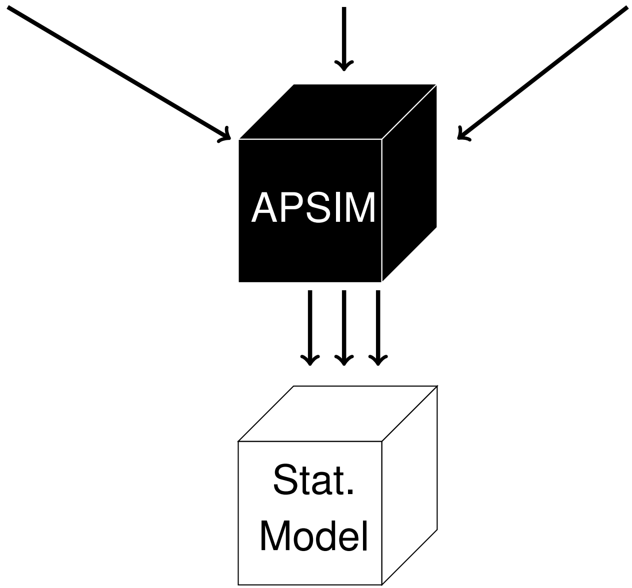
Traditional SA

$$\boldsymbol{X}_1 = (25, 0.041, 120, \ldots), \boldsymbol{X}_2 = (29, 0.038, 123, \ldots), \cdots, \boldsymbol{X}_N = (22, 0.045, 110, \ldots)$$



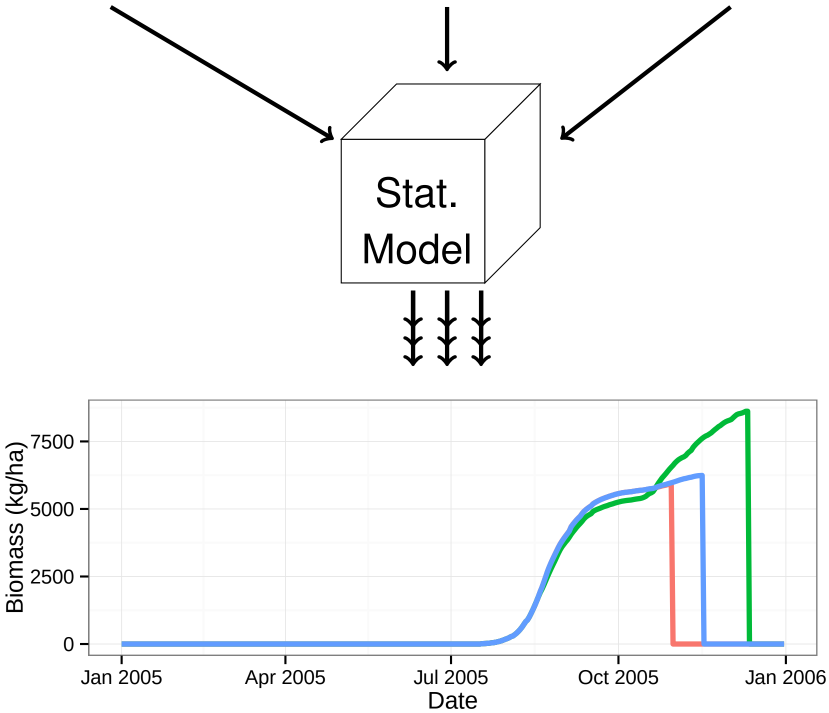
Emulators

$$\boldsymbol{X}_1 = (25, 0.041, 120, \dots), \boldsymbol{X}_2 = (29, 0.038, 123, \dots), \cdots, \boldsymbol{X}_{N/2} = (22, 0.045, 110, \dots)$$



Emulators

$$\boldsymbol{X}_1 = (25, 0.041, 120, \ldots), \boldsymbol{X}_2 = (29, 0.038, 123, \ldots), \cdots, \boldsymbol{X}_{N/2} = (22, 0.045, 110, \ldots)$$



GAM-based Emulator

- Y - univariate computer model output
- $\boldsymbol{X} = (X_1, \dots, X_p)$ - computer model inputs

- Emulate the computer model output $Y = f(\boldsymbol{X})$ with

$$\hat{f}(\boldsymbol{X}) = \hat{f}_0 + \sum_{i=1}^{p} \hat{f}_i(X_i) + \sum_{i=1}^{p-1} \sum_{j=i+1}^{p} \hat{f}_{ij}(X_i, X_j)$$

where

  - $\hat{f}_0$ is the estimated mean of Y
  - $\hat{f}_i(X_i)$ is a thin plate regression spline
  - $\hat{f}_{ij}(X_i, X_j)$ is a tensor product of the marginal smooths of X_i and X_j

GAM-based Emulator

- $x_{i1}, \ldots, x_{in}$ are generated values for X_i and $\boldsymbol{y} = (y_1, \ldots, y_n)$ the computer model output

- An estimate of S_i is given by $\hat{S}_i = \hat{V}_i / \mathrm{Var}(\boldsymbol{y})$ where

$$\hat{V}_i = \widehat{\mathrm{Var}}(\hat{f}_i) = \frac{1}{n-1} \sum_{j=1}^{n} \left[\hat{f}_i(x_{ij}) - \bar{\hat{f}}_i \right]^2$$

and $\bar{\hat{f}}_i = \sum_j \hat{f}_i(x_{ij})/n$

Multivariate Sensitivity Analysis

- Univariate SA methods can be extended to time series data
- Let $Y(t)$ denote computer model output for time $t = 1, \dots, T$
  1. Compute univariate sensitivity indices at each time step $S_i(t)$
  2. Summarize output over time and compute sensitivity indices for that summary, e.g. compute S_i for $\overline{Y} = \sum_t Y(t)/T$
  3. Choose a set of basis functions $\phi_k(t)$ and compute sensitivity indices for the basis function coefficients h_k

$$Y(t) - \overline{Y} = \sum_{k=1}^{T} h_k \phi_k(t)$$

Multivariate Sensitivity Analysis

- Univariate SA methods can be extended to time series data
- Let $Y(t)$ denote computer model output for time $t = 1, \ldots, T$

1. Compute univariate sensitivity indices at each time step $S_i(t)$
2. Summarize output over time and compute sensitivity indices for that summary, e.g. compute S_i for $\overline{Y} = \sum_t Y(t)/T$
3. Choose a set of basis functions $\phi_k(t)$ and compute sensitivity indices for the basis function coefficients h_k

$$Y(t) - \overline{Y} \approx \sum_{k=1}^{K} h_k \phi_k(t), \quad K << T$$

Emulating Spatio-Temporal Data

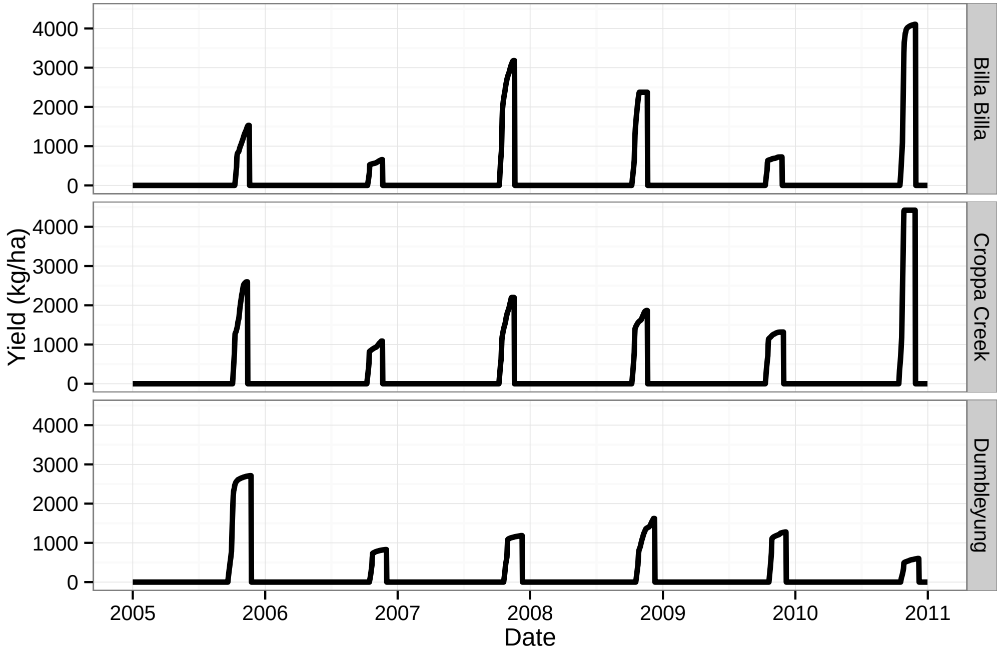
1. Run the computer model a reasonable number of times
2. Choose a few basis functions to reduce the dimensionality of the computer model runs
3. Use a low-dimensional GAM to emulate the coefficients of the chosen basis functions
4. Estimate the first-order and total sensitivity indices for each of the selected dimensions
5. Interpret your results

Application to APSIM

- Model daily **wheat yield** estimates from 2000 - 2010 in QLD, NSW and WA

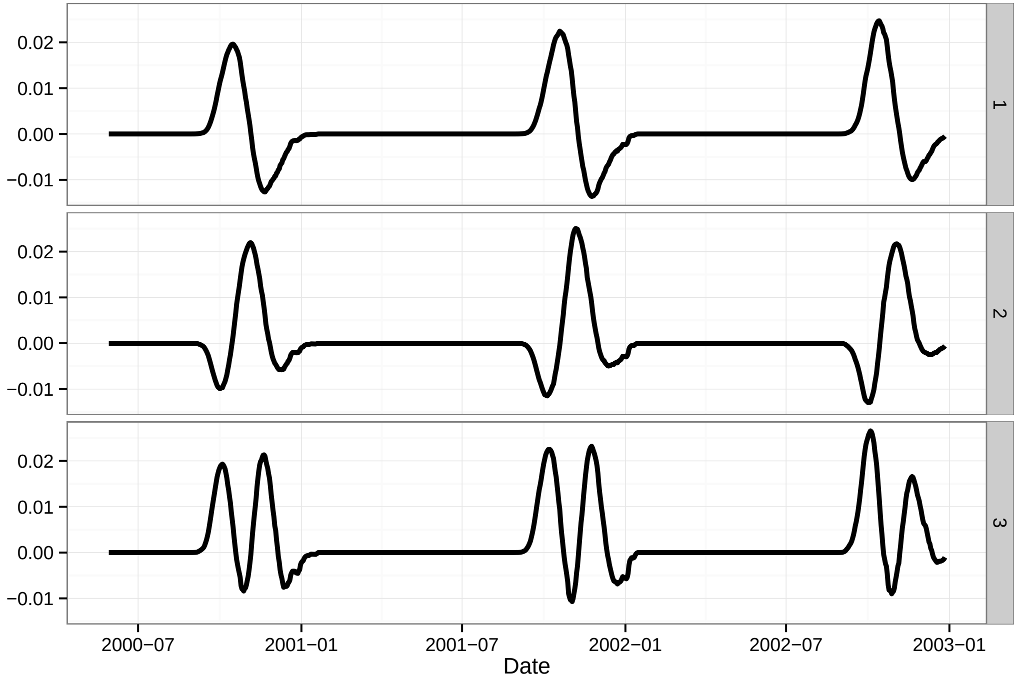
Step 1: Run Computer Model

- Model daily **wheat yield** estimates from 2000 - 2010 in QLD, NSW and WA



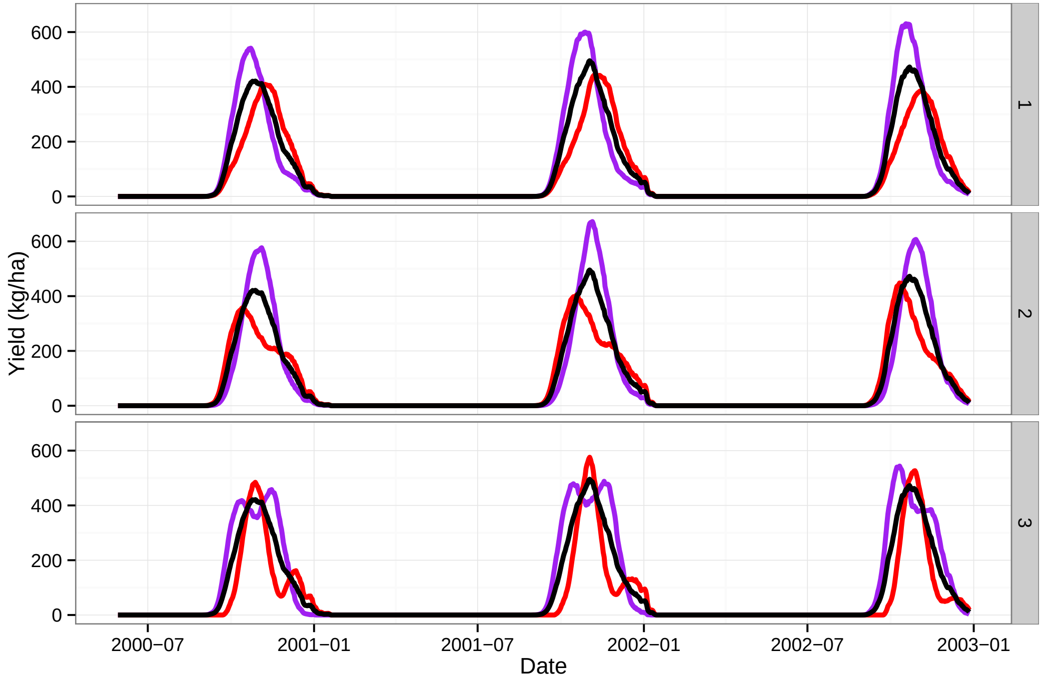
Step 2: Reduce Dimensionality

- Three principal components explain 32.1%, 24.5% and 8.0% of the variability

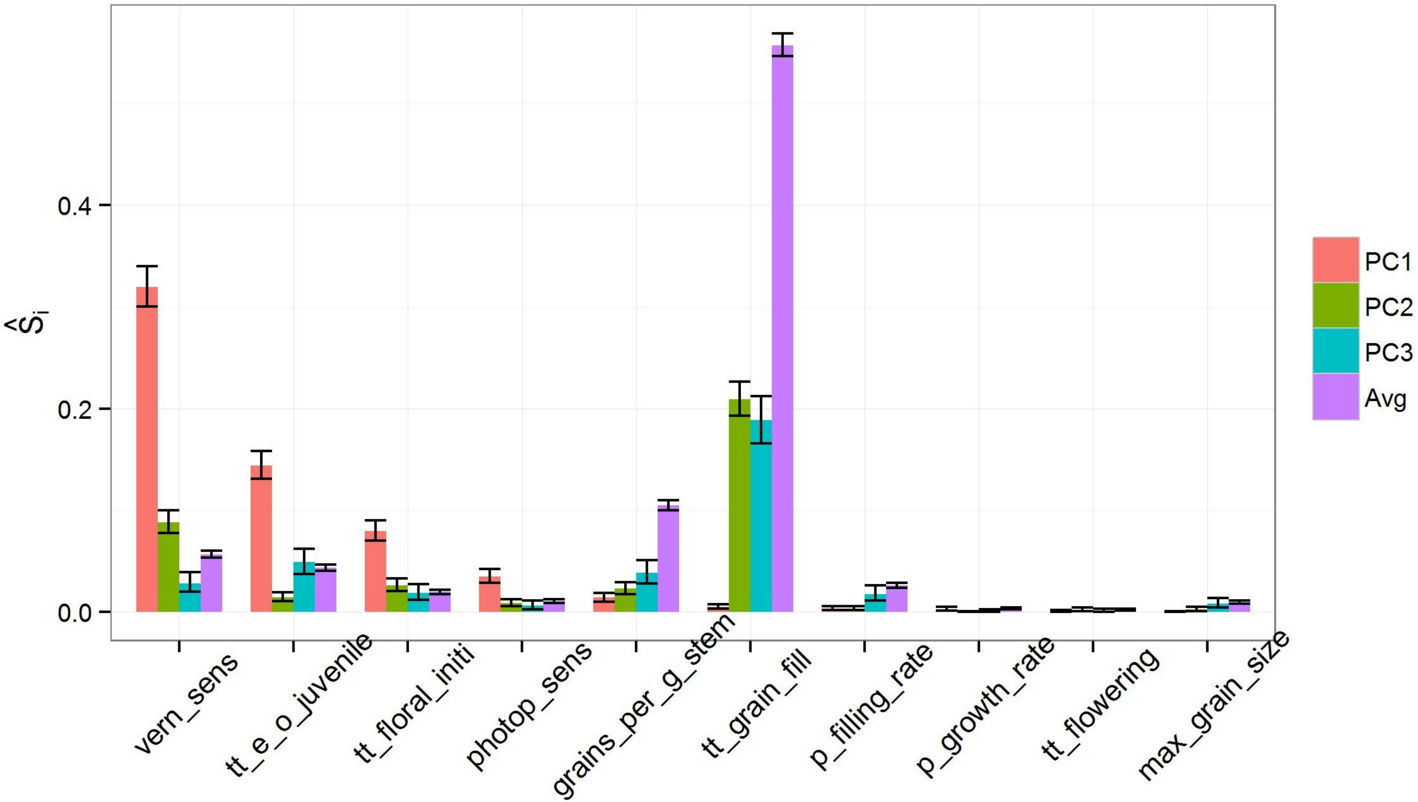


Step 2: Reduce Dimensionality

- Three principal components explain 32.1%, 24.5% and 8.0% of the variability



Steps 3 & 4: Fit GAM & Compute $\hat{S}_i$



Step 5: Interpret Results

- Uncertainty in mode location and height are the first and second largest sources of variability in wheat yield estimates, respectively
- Variability in mode location is due mainly to variability in length of vegetative growth period (“vern_sens”) and the length of thermal time between the juvenile phase and floral initiation (“tt_e_o_juvenile”)
- Variability in yield magnitude is due mainly to time allowed for crop to fill grains (“tt_grain_fill”)

Take aways

- A simple low-dimensional GAM is an efficient emulator for many computer models and is easy to program (`mgcv` package in R)

- SA with univariate summaries of functional data can give misleading* results

- The uncertainty in a computer model and the uncertainty in the response is not always the same

Ongoing work

- Principal components are often difficult to interpret

- Proposed emulation method can give biased total sensitivity index estimates if higher-order interactions are present

- GAMs can require a lot of data and time if the number of inputs is large

Thank You

CSIRO Data Analytics

Bryan Stanfill

t +61 7 3833 5727

e Bryan.Stanfill@csiro.au

w http://stanfill.github.io/

This research is supported by the Science and Industry Endowment Fund.

SCIENCE AND INDUSTRY ENDOWMENT FUND