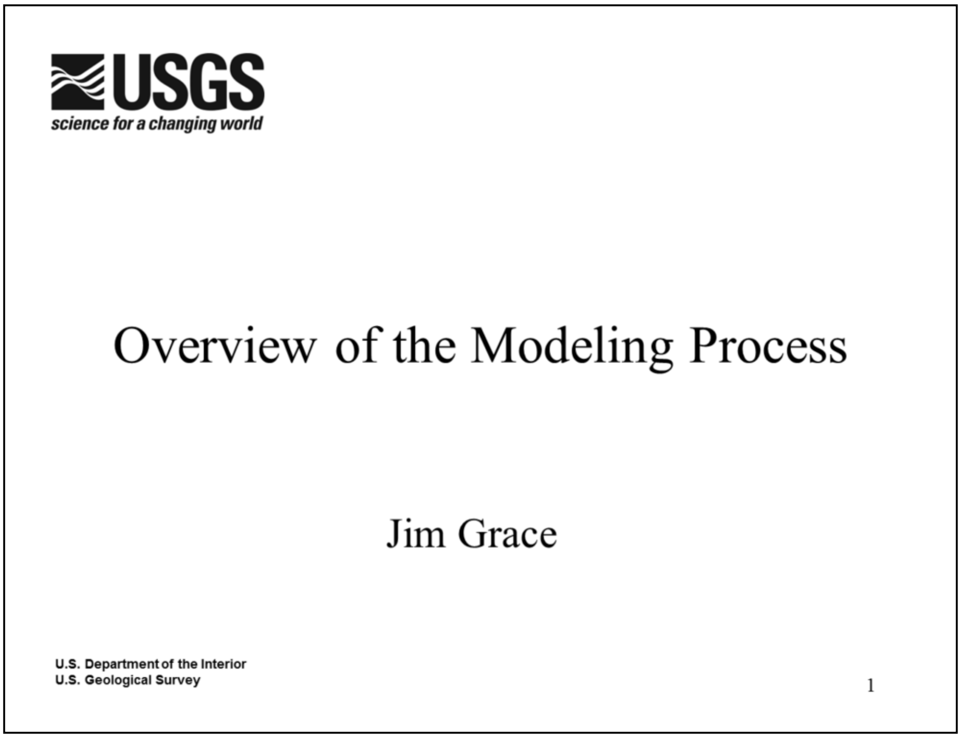


Overview of the Modeling Process

Jim Grace

U.S. Department of the Interior
U.S. Geological Survey

1

This very brief module provides a very general set of points about the overall modeling process.

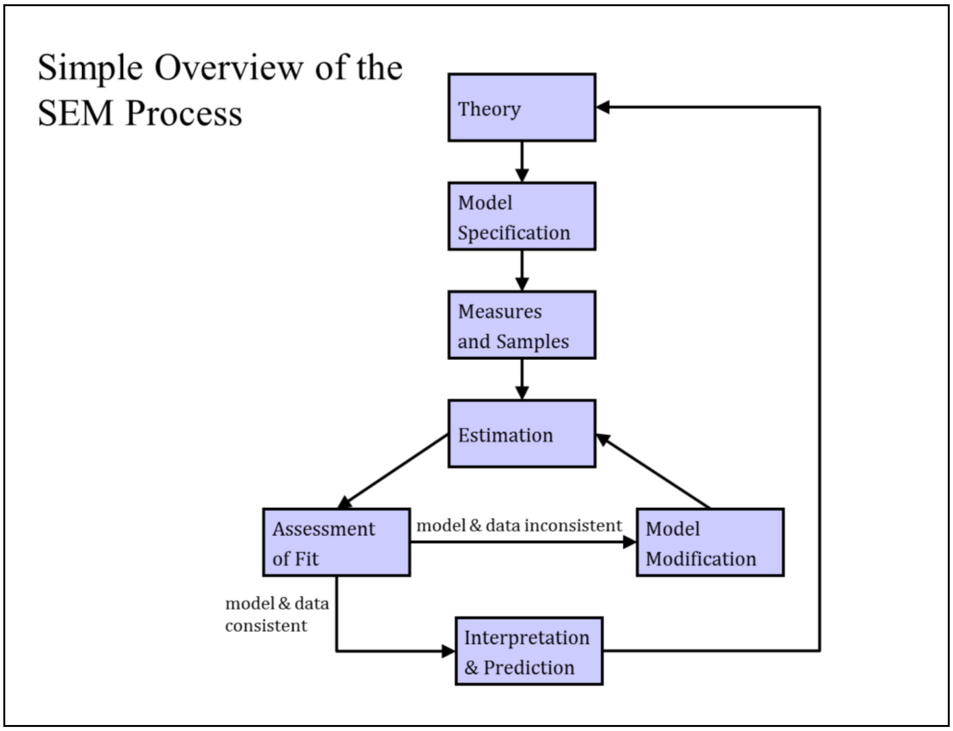
An appropriate citation for the material in this tutorial is

Grace, J.B., Anderson, T.M., Olff, H., and Scheiner, S.M. 2010. On the specification of structural equation models for ecological systems. *Ecological Monographs* 80:67-87.

Notes: IP-056512; Support provided by the USGS Climate & Land Use R&D and Ecosystems Programs. I would like to acknowledge formal review of this material by Jesse Miller and Phil Hahn, University of Wisconsin. Many helpful informal comments have contributed to the final version of this presentation. The use of trade names is for descriptive purposes only and does not imply endorsement by the U.S. Government.

Last revised 17.02.05.

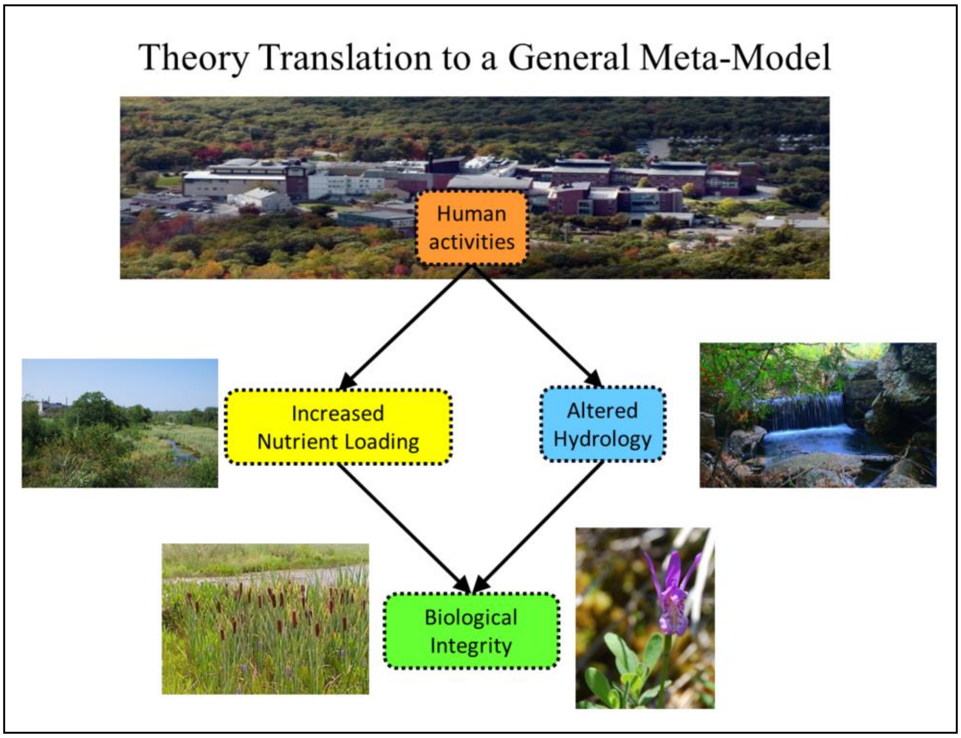
Source: https://www.usgs.gov/centers/wetland-and-aquatic-research-center/science/quantitative-analysis-using-structural-equation



Simple Overview of the SEM Process

SEM is a process designed to lead to scientifically interpretable models. It relies on sequential learning and a general multi-step process to build confident knowledge.

In SEM, we first translate our ideas into models, test those models, modify our models if need be, and then use that knowledge to inform where we start with the next study.



Theory Translation to a General Meta-Model

Human activities

Increased Nutrient Loading

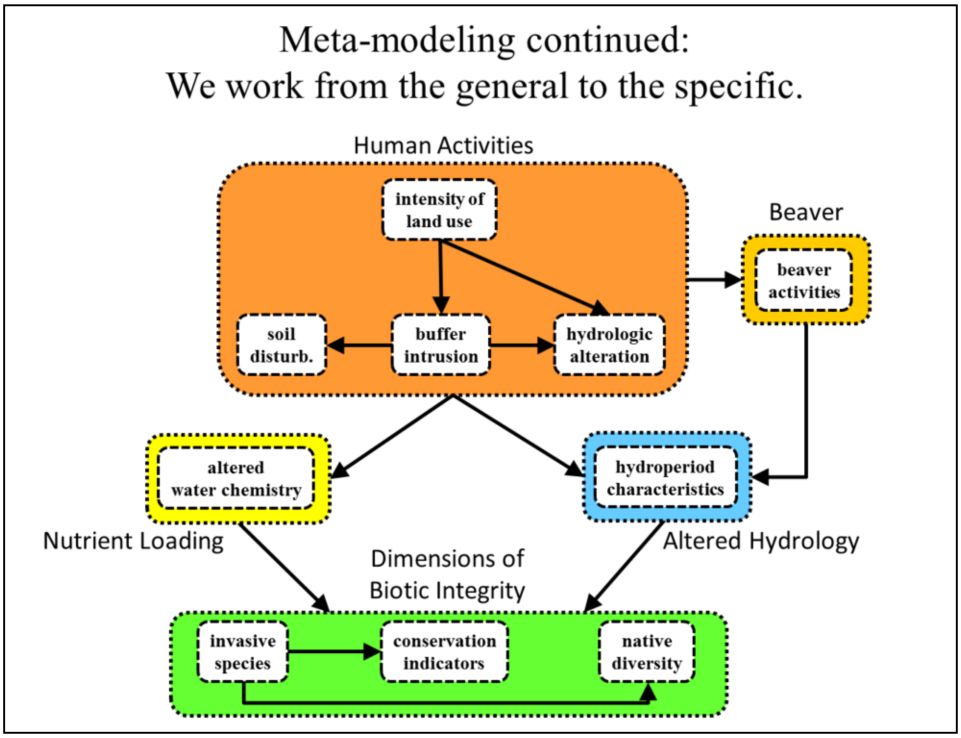
Altered Hydrology

Biological Integrity

Theory translation should be as explicit as possible. Most modeling exercises have historically translated theory to models using an informal approach and have often omitted many important steps.

In Grace et al. 2010. Ecological Monographs, we propose meta-modeling as a process for defining the family of more specific hypotheses we will will examine.

In meta-modeling, we start with a general view of the problem. Background (a priori) knowledge is often only defined at the linguistic level. We strive to show how our model specifics (shown in later slides) relate to the general a priori knowledge. Here I show the a very general meta-model that represents the most general logic for the analysis.



Meta-modeling continued:
We work from the general to the specific.

Human Activities

intensity of land use

soil disturb.

buffer intrusion

hydrologic alteration

Beaver

beaver activities

altered water chemistry

Nutrient Loading

hydroperiod characteristics

Altered Hydrology

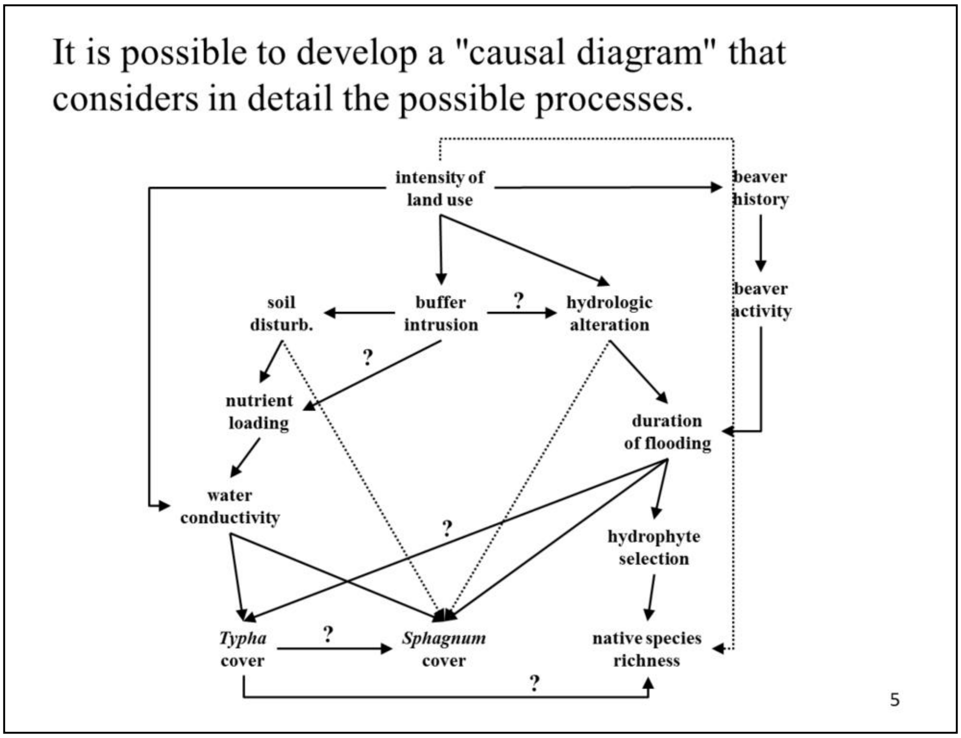
Dimensions of Biotic Integrity

invasive species

conservation indicators

native diversity

Considering the available data, we can build out our theoretical (a priori) expectations. At this point in the process we remain conceptual and do not yet consider statistical issues like linear relations or response variable distributions.

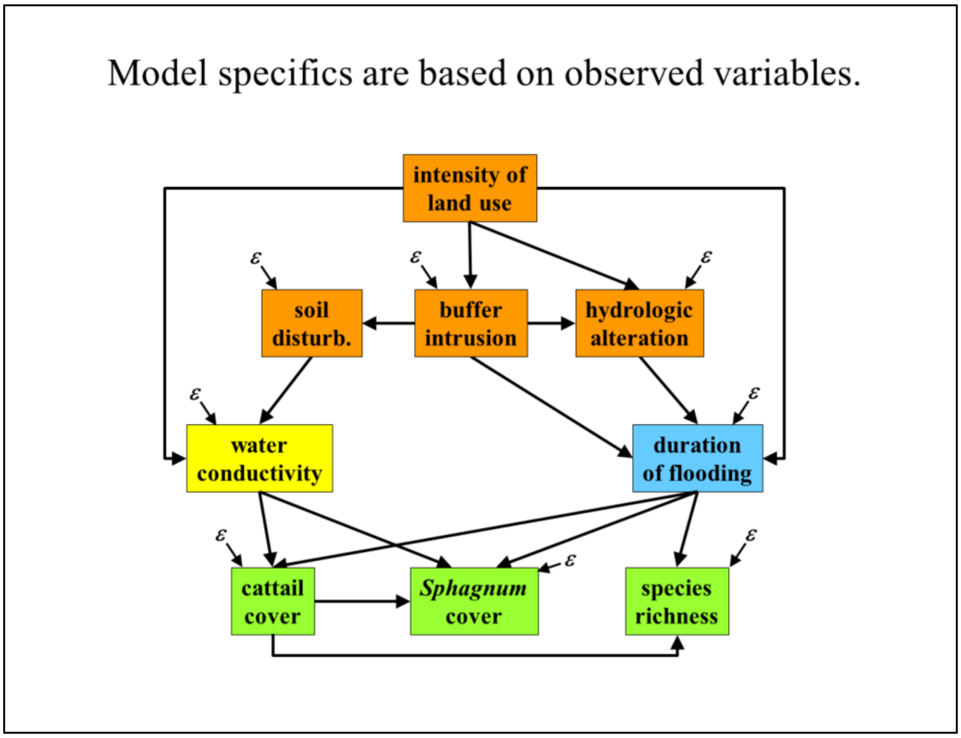


It is possible to develop a "causal diagram" that considers in detail the possible processes.

The idea of a causal diagram comes from Judea Pearl's work on causality.

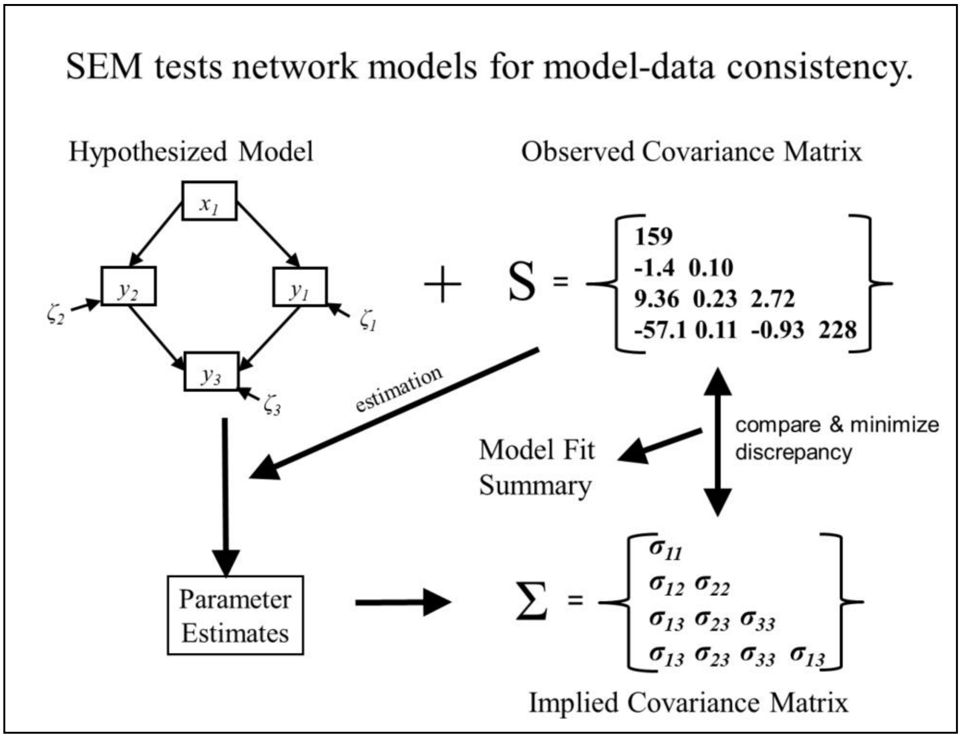
Pearl, J. 1995. Causal diagrams for empirical research. Biometrika 82:669–710.

Causal diagrams are meant to serve as a template for modeling. The can include variables that we do not have measurements for and are meant to represent more deeply the mechanisms of interest. Pearl also means for these diagrams to help us identify the observations that would be needed to estimate various parameters using limited data.



Model specifics are based on observed variables.

The next major step in the process to fully specify our models. Now we have to work with the variables for which we have observations. In this case, “water conductivity” serves as our indicator of nutrient loading in these very soft-water systems. SEM procedures allow us to test whether the data are consistent with the implications of this architecture. This means we can test for whether there are any linkages we have omitted or any linkages that cannot be supported by the data.

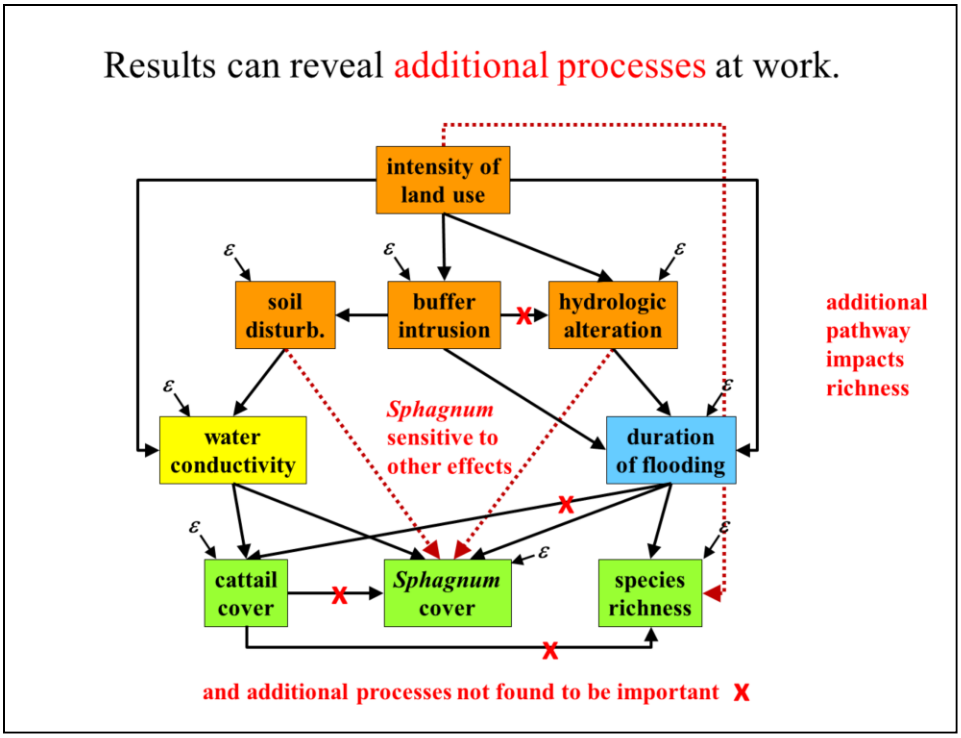


SEM tests network models for model-data consistency.

$$S = \begin{Bmatrix} 159 \\ -1.4\ \ 0.10 \\ 9.36\ \ 0.23\ \ 2.72 \\ -57.1\ 0.11\ \ -0.93\ \ 228 \end{Bmatrix}$$

$$\Sigma = \begin{Bmatrix} \sigma_{11} \\ \sigma_{12}\ \sigma_{22} \\ \sigma_{13}\ \sigma_{23}\ \sigma_{33} \\ \sigma_{13}\ \sigma_{23}\ \sigma_{33}\ \ \sigma_{13} \end{Bmatrix}$$

Here is a cartoon to represent the sequence of steps in traditional global estimation. For alternative estimation approaches, consult the module on Estimation.



Results can reveal additional processes at work.

additional pathway impacts richness

Sphagnum sensitive to other effects

and additional processes not found to be important X

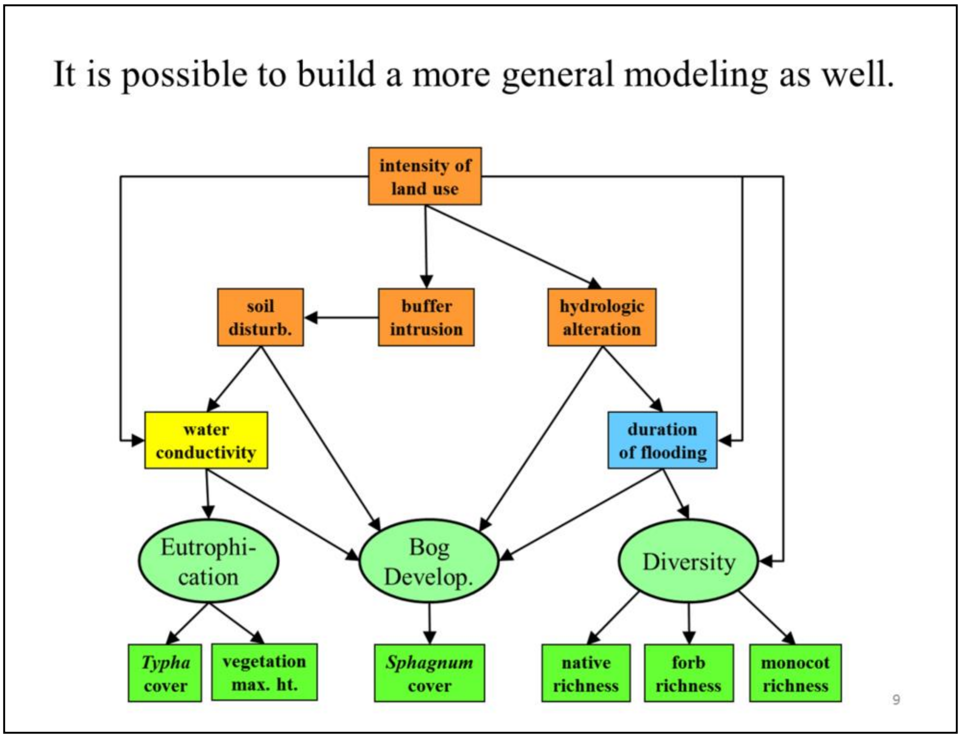
Once tested with data, we did indeed find evidence for a number of linkages (and associated processes) that we did not initially anticipate.

(1) It seems that species richness is lower where there is higher landuse intensity for some additional reason beyond altered duration of flooding.

(2) Sphagnum moss communities, a characteristic element of the natural system, seems to have been impacted through some additional methods shown as direct paths from soil disturbance and hydrologic alteration.

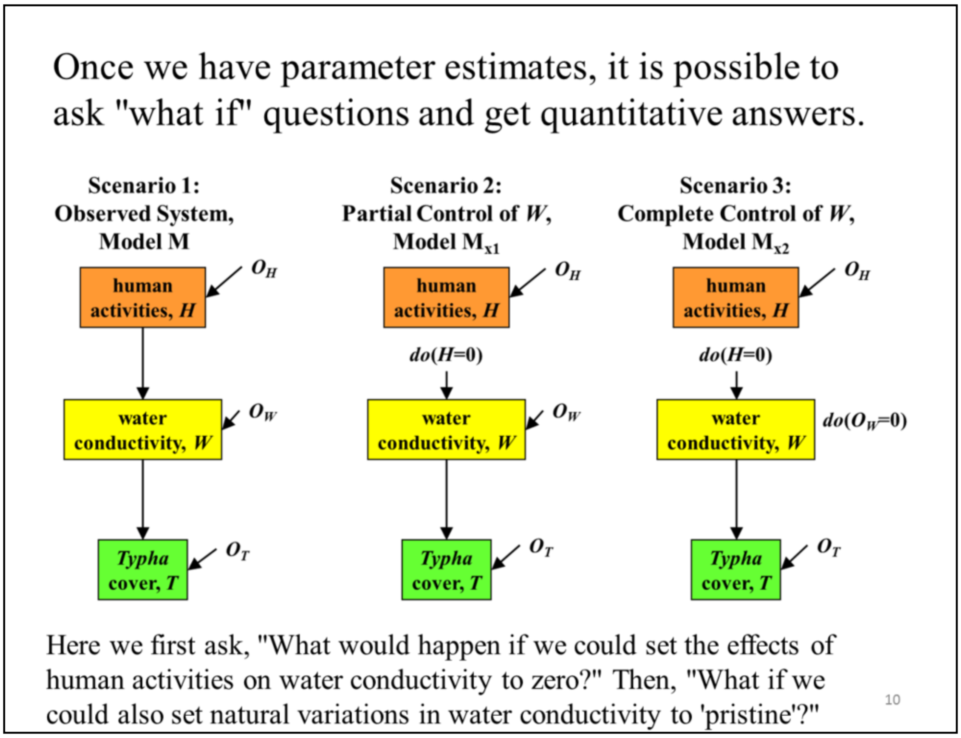
(3) Surprisingly, no real indication in this sample that impounding (increasing the duration of flooding) affects cattail invasions.

(4) So far, cattails don’t seem to have had a detectable impact on Sphagnum or native species richness, though we certainly expect such effects could occur.



It is possible to build a more general modeling as well.

This slide presents Figure 13 from our 2012 Ecosphere paper. Here we show an alternative structural equation model that uses latent variables to represent the generalized system responses (Eutrophication, Bog Development, and Diversity) that lead to the observed biological metric values. Error variables are not shown for simplicity.

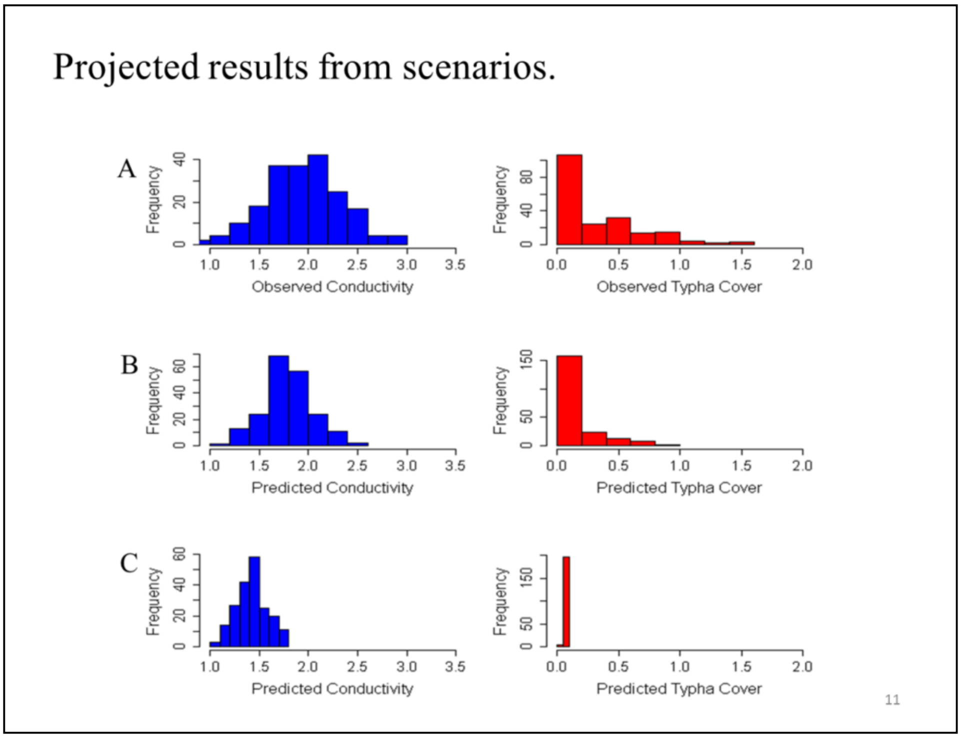


Once we have parameter estimates, it is possible to ask "what if" questions and get quantitative answers.

Here we first ask, "What would happen if we could set the effects of human activities on water conductivity to zero?" Then, "What if we could also set natural variations in water conductivity to 'pristine'?"

10

This figure is Figure 11 from Grace et al. 2012 Ecosphere. Queries about predicted effects of interventions on water conductivities and cattail abundance. Scenario 1 is status quo; Scenario 2 is elimination of buffer intrusion and soil disturbance; Scenario 3 is reduction of water conductivity to reference conditions. The “O” variables refer to unknown and unspecified causes of variation (sometimes referred to as U for unspecified). The operator “*do*(*H=0*)” refers to reducing the values of land use and soil disturbance on conductivity to 0. The operator “do(O_W=0)” refers to reducing the value of other (unknown) factors on conductivity to 0.



Projected results from scenarios.

Figure 12, Grace et al. 2012 Ecosphere. A. Observed distribution of values of cattail abundance. B. Predicted distribution for case where effects of human activities are eliminated (Model M_{x1} in Fig. 11). C. Predicted distribution for case where conductivity is controlled (Model M_{x2}). Consult the original reference for more details.

Supporting Literature

James B. Grace

Structural Equation Modeling and Natural Systems

Cambridge

Grace (2006)
Chapter 10

CONCEPTS & SYNTHESIS

EMPHASIZING NEW IDEAS TO STIMULATE RESEARCH IN ECOLOGY

Ecological Monographs, 80(1), 2010, pp. 67–87
© 2010 by the Ecological Society of America

On the specification of structural equation models for ecological systems

James B. Grace,[1,4] T. Michael Anderson,[2,5] Han Olff,[2] and Samuel M. Scheiner[3]

[1] *U.S. Geological Survey, National Wetlands Research Center, 700 Cajundome Boulevard, Lafayette, Louisiana 70506 USA*
[2] *Community and Conservation Ecology Group, Centre for Ecological and Evolutionary Studies, University of Groningen, P.O. Box 14, 9750 AA Haren, The Netherlands*
[3] *Division of Environmental Biology, National Science Foundation, Arlington, Virginia 22230 USA*

esa ECOSPHERE

SYNTHESIS & INTEGRATION

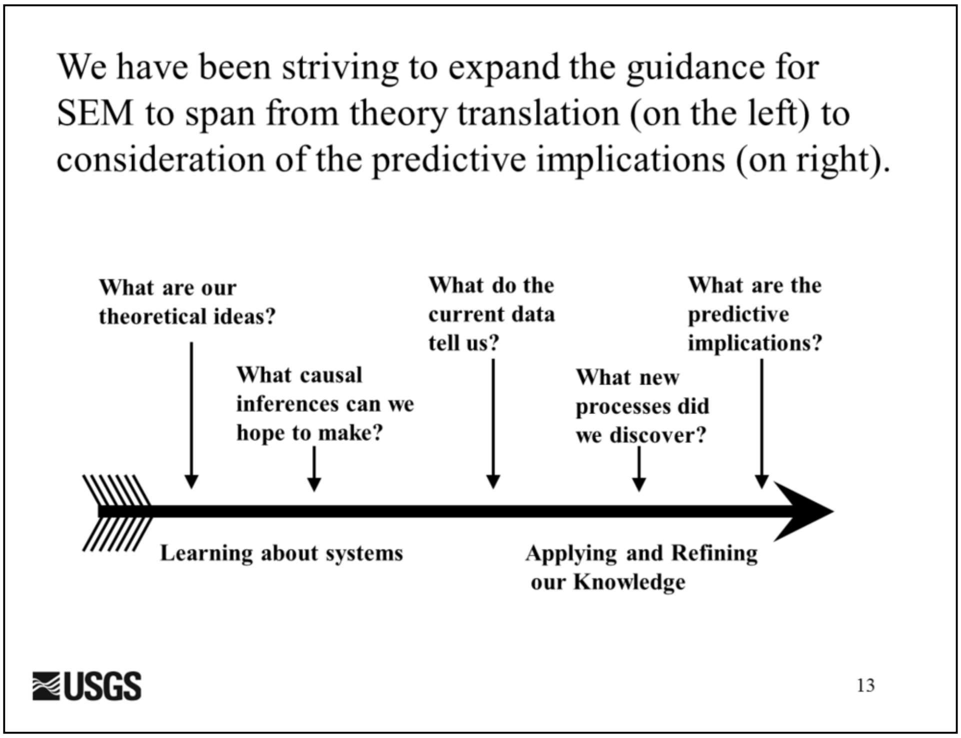
Guidelines for a graph-theoretic implementation of structural equation modeling

James B. Grace,[1,†] Donald R. Schoolmaster Jr.,[2] Glenn R. Guntenspergen,[3] Amanda M. Little,[4] Brian R. Mitchell,[5] Kathryn M. Miller,[6] and E. William Schweiger[7]

[1] *U.S. Geological Survey, National Wetlands Research Center, 700 Cajundome Boulevard, Lafayette, Louisiana 70506 USA*
[2] *Five Rivers Services, LLC at the U.S. Geological Survey, National Wetlands Research Center, 700 Cajundome Boulevard, Lafayette, Louisiana 70506 USA*
[3] *U.S. Geological Survey, Patuxent Wildlife Research Center, Laurel, Maryland 20708 USA*
[4] *Biology Department, University of Wisconsin-Stout, Menomonie, Wisconsin 54751 USA*
[5] *National Park Service, Northeast Temperate Network, Woodstock, Vermont 05091 USA*
[6] *National Park Service, Northeast Temperate Network, Acadia National Park, Bar Harbor, Maine 04609 USA*
[7] *National Park Service, Rocky Mountain Network, Fort Collins, Colorado 80525 USA*

USGS

I have published three different treatments of the SEM workflow process. In my book, I walk through an example to illustrate the intent of sequential learning about a problem. In 2010, we expounded on model specification choices and the grounding needed for decisions in the field of ecology. In 2012, we proposed a 3rd-generation implementation for SEM and some additional steps in the process.



We have been striving to expand the guidance for SEM to span from theory translation (on the left) to consideration of the predictive implications (on right).

What are our theoretical ideas?

What causal inferences can we hope to make?

What do the current data tell us?

What new processes did we discover?

What are the predictive implications?

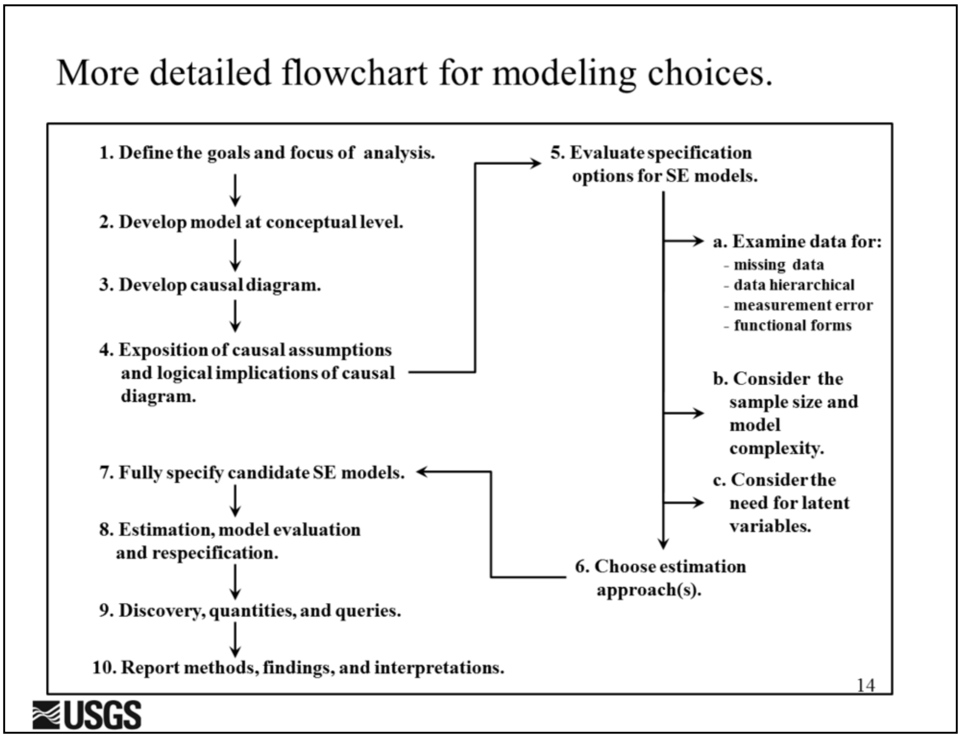
Learning about systems

Applying and Refining our Knowledge

13

From a science perspective, one of our objectives has been to expand the advice given, as well as the procedures that link questions to answers. There has been a substantial gap in the literature on SEM dealing with the ends of the process. On the front-end, how do we formally translate theoretical ideas into models in a “revealed” fashion. On the back-end, there are many possible uses for our hard-earned parameter estimates. This potential is largely untapped because of a lack of attention by SEMers to issues that are bread-and-butter of “modelers”.

What we aspire to is a comprehensive system for quantitatively examining general theoretical ideas.



More detailed flowchart for modeling choices.

1. Define the goals and focus of analysis.

2. Develop model at conceptual level.

3. Develop causal diagram.

4. Exposition of causal assumptions and logical implications of causal diagram.

5. Evaluate specification options for SE models.

a. Examine data for:
- missing data
- data hierarchical
- measurement error
- functional forms

b. Consider the sample size and model complexity.

c. Consider the need for latent variables.

6. Choose estimation approach(s).

7. Fully specify candidate SE models.

8. Estimation, model evaluation and respecification.

9. Discovery, quantities, and queries.

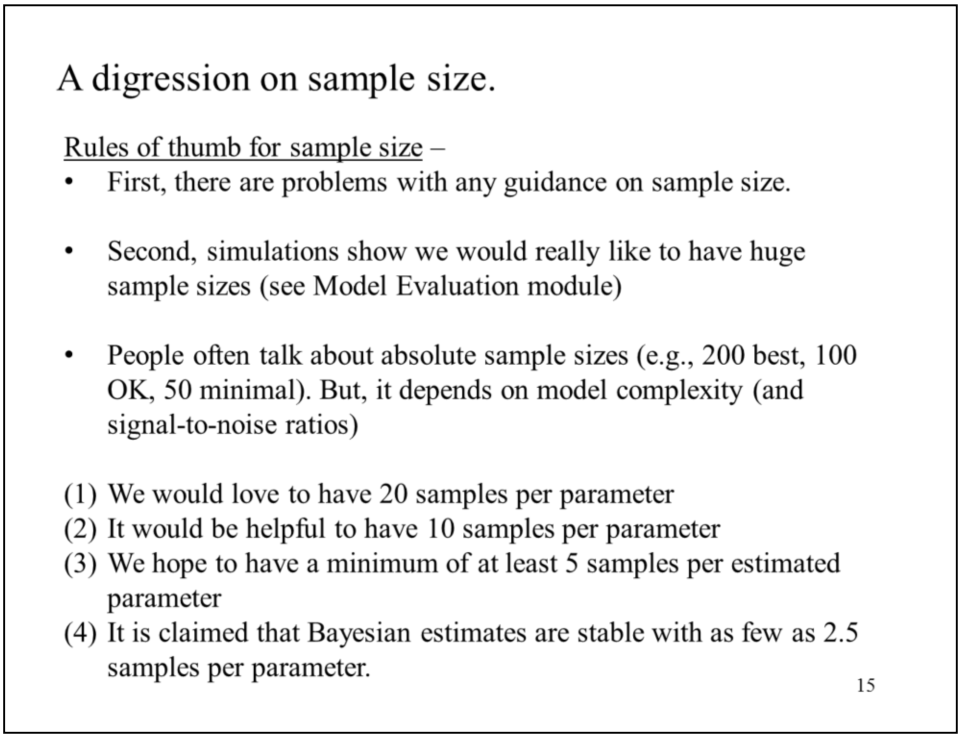
10. Report methods, findings, and interpretations.

14

USGS

These are the guidelines given in the Ecosphere paper. They are also elaborated on in a book chapter*.

*Grace, J.B., Scheiner, S.M., Schoolmaster, D.R. Jr. 2015. Structural equation modeling: building and evaluating causal models. Chapter 8 In: Fox, G.A., Negrete-Yanlelevich, S., and Sosa, V.J. (eds.) *Ecological Statistics: From Principles to Applications*. Oxford University Press. (accepted and in production)



A digression on sample size.

Rules of thumb for sample size –

- First, there are problems with any guidance on sample size.
- Second, simulations show we would really like to have huge sample sizes (see Model Evaluation module)
- People often talk about absolute sample sizes (e.g., 200 best, 100 OK, 50 minimal). But, it depends on model complexity (and signal-to-noise ratios)

(1) We would love to have 20 samples per parameter
(2) It would be helpful to have 10 samples per parameter
(3) We hope to have a minimum of at least 5 samples per estimated parameter
(4) It is claimed that Bayesian estimates are stable with as few as 2.5 samples per parameter.

Here is just a very little bit about sample size giving some common rules of thumb. A general reference for the topic is

Kline, RB 2015. Principles and Practice of Structural Equation Modeling. Guilford Press.