10-9-2015

Prediction: The Quintessential Model Validation Test

Wayne Wakeland
Portland State University, wakeland@pdx.edu

Prediction--the Quintessential Policy Model Validation Test

Wayne Wakeland

Systems Science Seminar Presenation
10/9/15

Assertion

- Models must, of course, be well suited to their intended application
- Thus, models for evaluating policies must be able to "predict" how the system is likely to respond to alternative policies
  - To a useful degree, and over a relevant time period
- One must, therefore, compare model predictions to what actually happens
- As recommended in the SD literature
  - But is rarely demonstrated

Must one Wait for Future to Unfold?

- It might be possible, for example, to blind oneself to the recent past and the use distant past to predict the more recent past
  - Concern is whether modeler is truly blind
    - Even a glance at a graph of recent outcomes could introduce subjective bias
- Another approach could be to use an algorithm for model calibration
  - Algorithms much less susceptible to subjective bias
- Predicting unknown future would be the most compelling test

Background

- Model testing has received considerable attention in SD literature
  - Key resources: Forrester and Senge 1980, Barlas 1996, Coyle and Exelby 2000, Sterman 2000, Olivia 2003 ,Saysel and Barlas 2006, Martis 2006 ,Groesser and Schwaninger 2012, and many more
  - Predictive capability discussed some detail, but few examples are provided
- Model testing was often referred to as verification and validation
  - Authors have tended to avoid the word “validation” in order to avoid confusion with concept of statistical validity
    - Or the implication that SD models can be declared valid or invalid by running a set of tests
  - Emphasis is on rigorous and thorough testing processes, and establishing a model’s domain or boundary of applicability

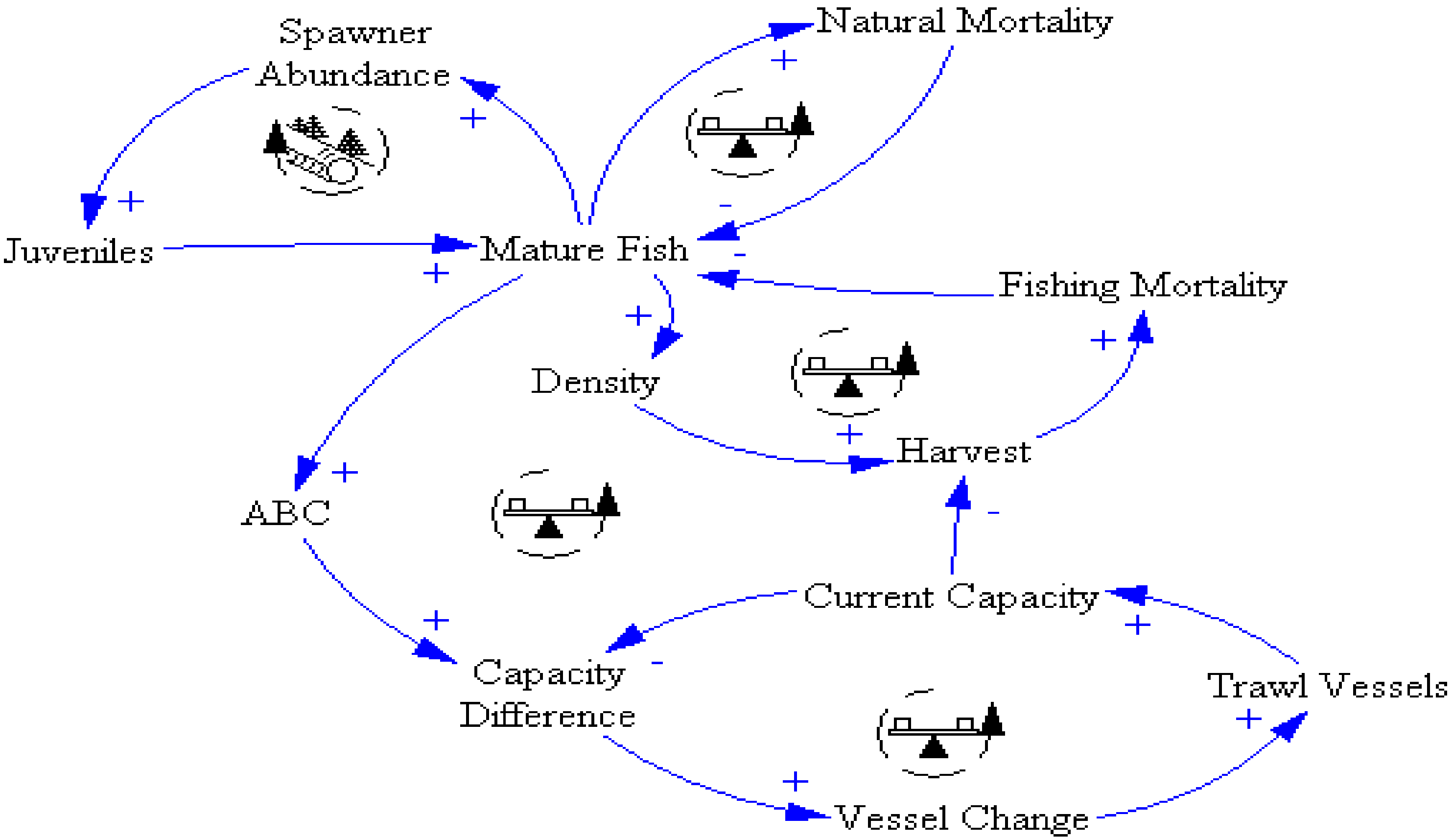
Methods

- Revisit three SD policy / prescriptive models to determine accuracy of their predictions
  - In each case, model emphasized calibration of model against historical reference behavior
- Further, to examine underlying causes of prediction failures

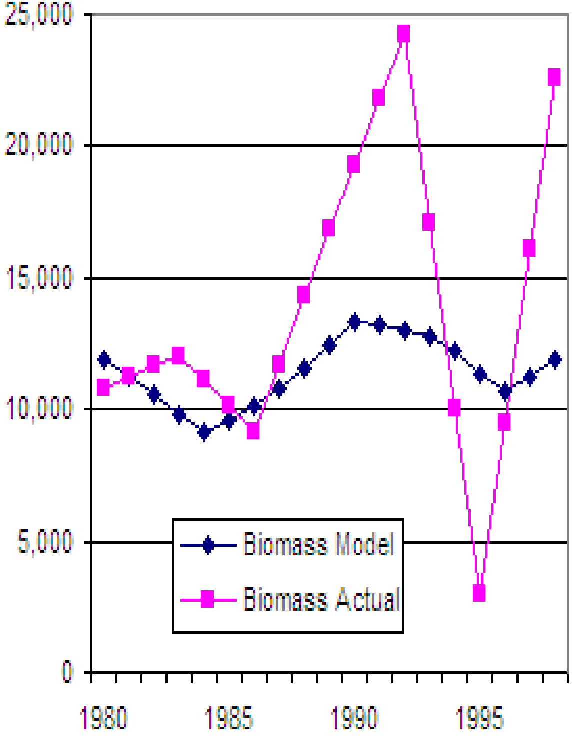
Case 1: Fishery Regulation

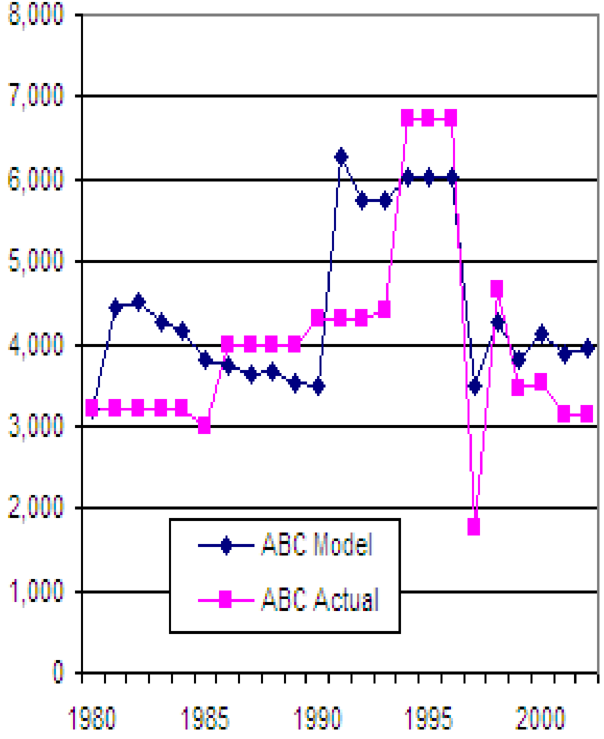
- Stopping the decline of fish populations is very challenging
  - Rockfish landings were down nearly 80% and catch limits had been reduced by 78%-89%
  - West Coast ground fish fisheries were declared a federal disaster in 2000
- Likely due to ineffective natural resource management and short-term policies
  - Leading to a larger fishing fleet than could be supported
- Applications of SD to fisheries management are plentiful
  - Ruth and Lindholm 1996, Holland and Brazee 1996, Dudley and Soderquist 1999, Ford 1999, van den Belt 1999, Dudley 2003, Jentoft 2003, Moxnes 1998, 2000, 2004, 2005, Brekke and Moxnes 2003), Wakeland, et al 2003, Wakeland 2007

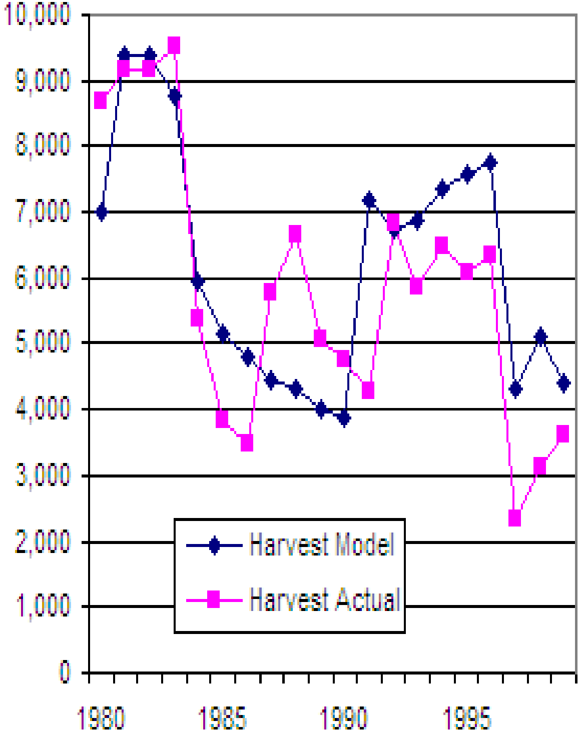
Case 1: High-level CLD for Fisheries



Case 1: Model Calculations vs. Reference Data







Biomass

Acceptable Catch

Harvest

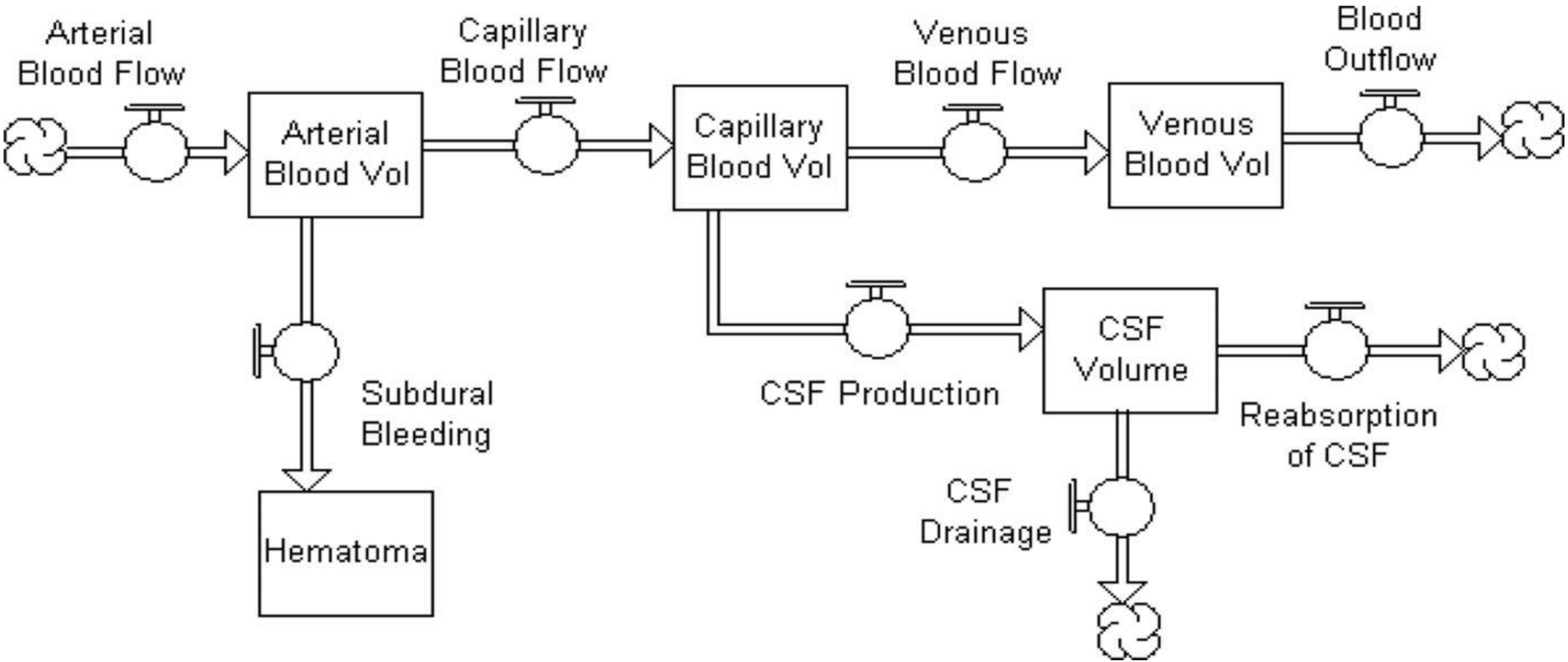
Case 2: Intracranial Pressure (ICP) Prediction

- Traumatic brain injury remains leading cause of death and disability in children
  - 30+% death rate for severe pedatric TBI
- Many sophisticated computer models have been created
- Parameters are typically estimated by calibrating models to fit patient-specific clinical data
  - Ursino and Lodi 1997, Ursino and Magosso 2001, Wakeland et al. 2005, Hu et al. 2007
  - Excellent results reported by Ursino and colleagues 2000
- Wakeland et al. 2009 was the first study to report actual prediction accuracy
  - Some studies refer to model calculations as predictions even though the study aim was to match ("predict") reference data
    - Ursino, Minassian, Lodi et al. 2000

Case 2: Data Collection

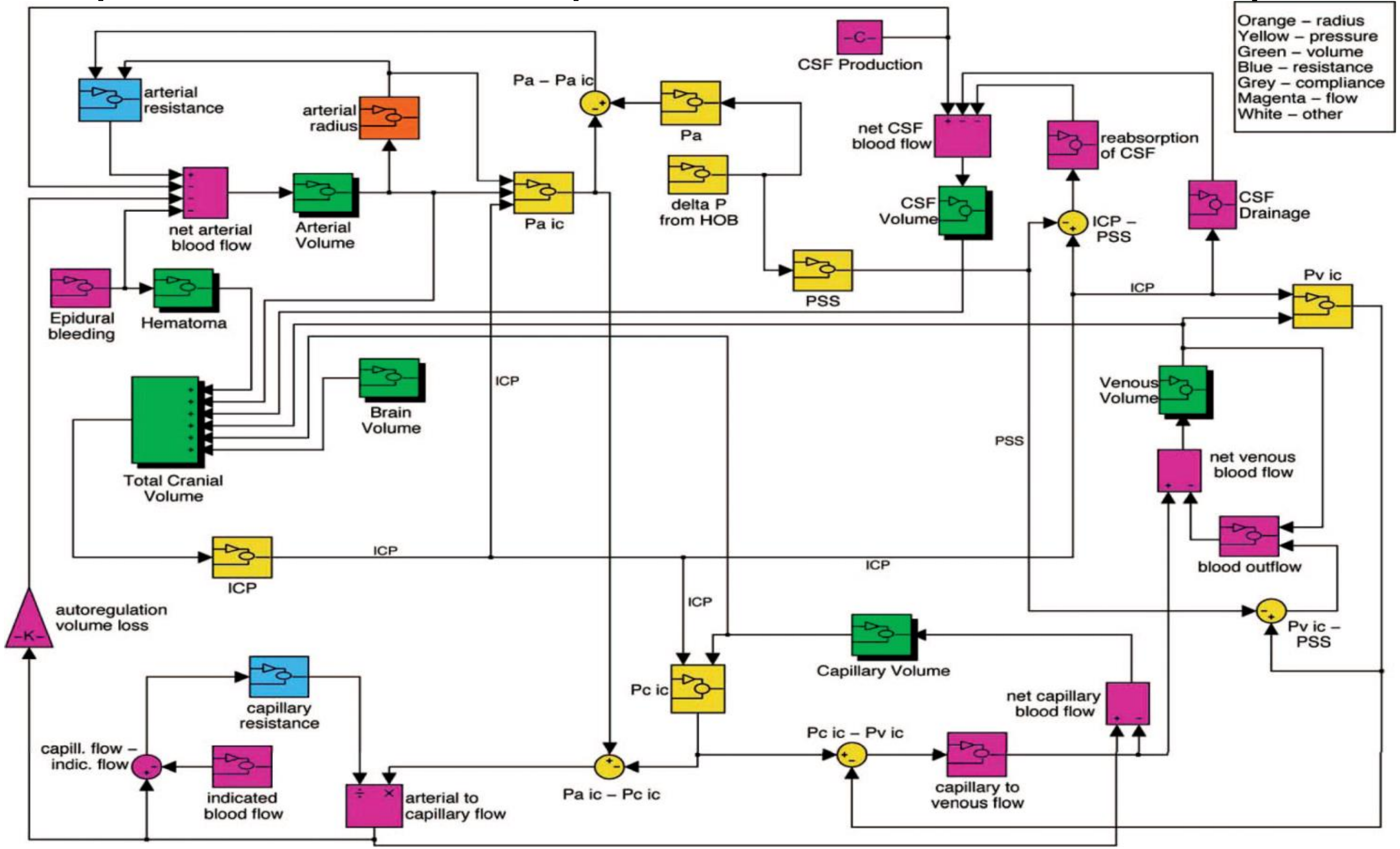
- Patients given mild [IRB-approved] physiological challenges to estimate their state of autoregulation
  - Changing the head of bed between 0 and 30 degrees
  - Changing respiration rate to create mild hyper-ventilation and mild hypo-ventilation
- Patient ICP response carefully measured and recorded
- Goal: determine if patient-specific models could predict patient ICP response to interventions
  - And, ultimately, to use them to evaluate alternative treatments beforehand "in silico"

Case2: Primary Stocks & Flows in ICP Dynamic Model

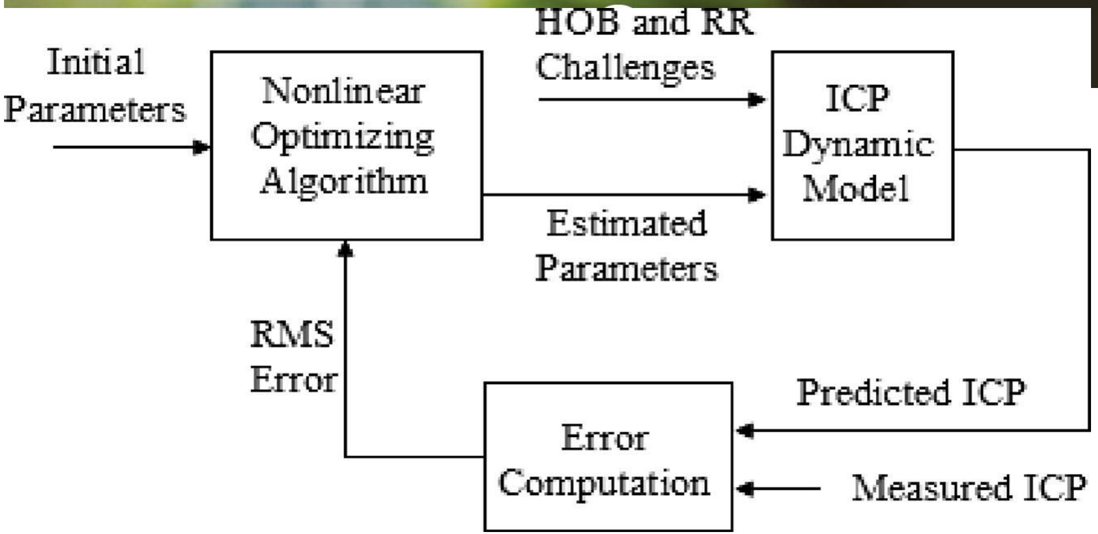


Case 2: ICP Model

(developed in STELLA and ported to Simulink for computation)

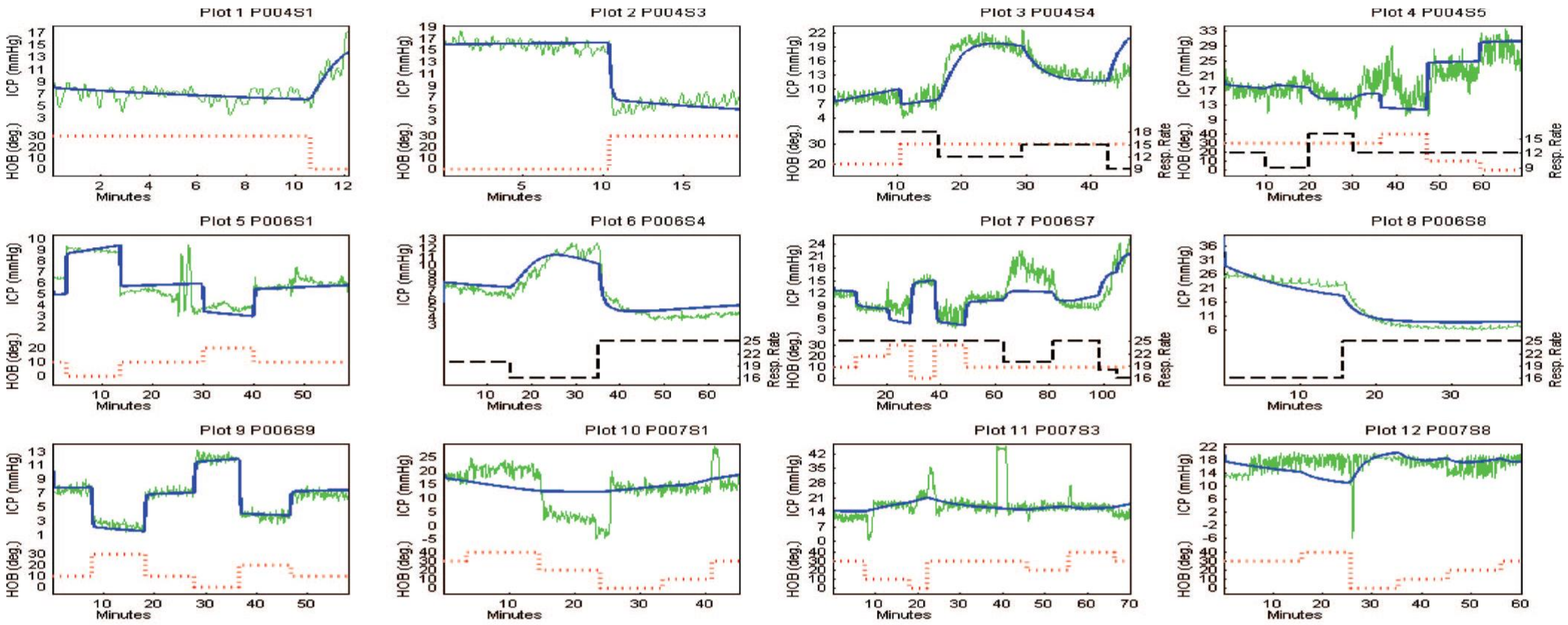


Case2: Parameter Estimation Process to Create Patient-specific models

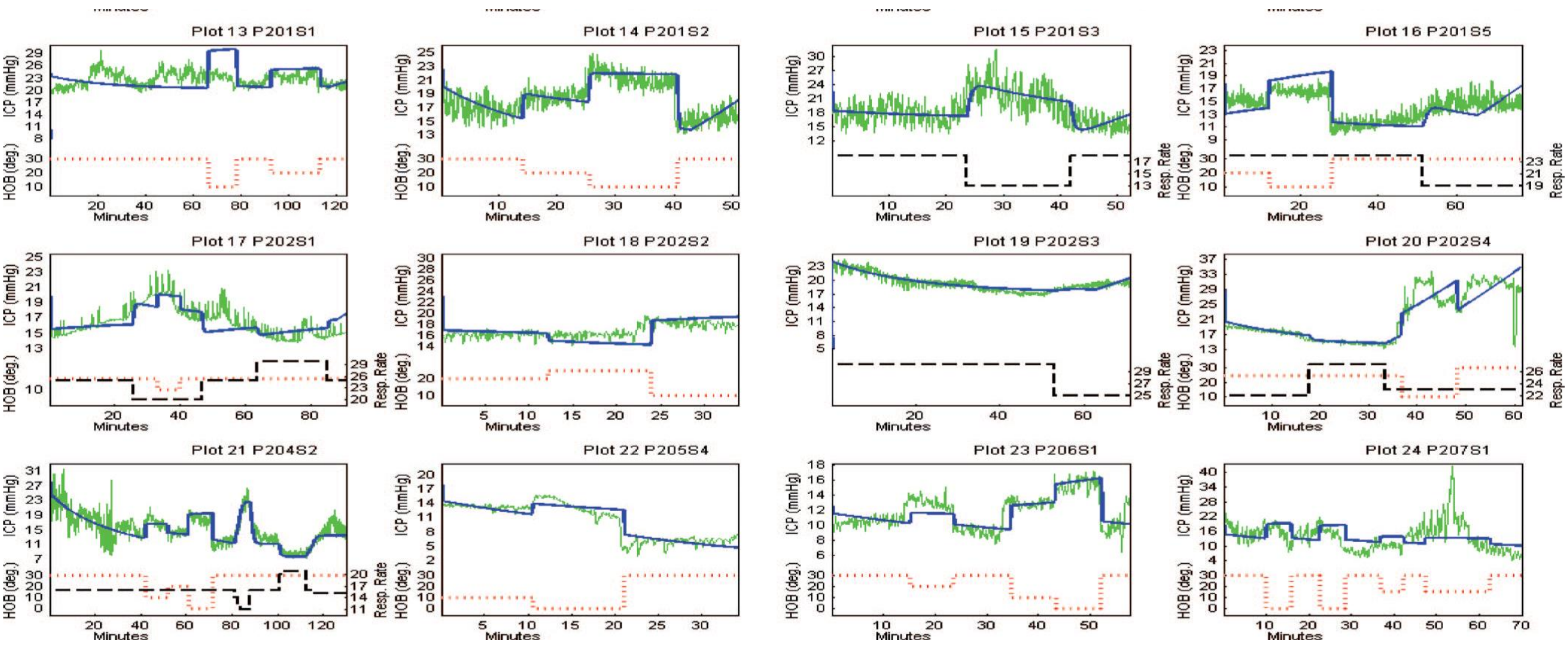


- **Parameters Estimated**
  - Autoregulation factor (smooth muscle compliance effect)
  - Basal cranial volume -- CSF drainage rate -- Hematoma increase rate
  - Δ pressure time constant (a smoothing parameter associated with HOB elevation change)
  - ETCO2 time constant (a smoothing parameter associated with RR changes)
  - Smooth muscle gain (a multiplicative factor related to the impact of smooth muscle tension)
  - Systemic venous pressure -- “Baseline” ICP -- Pressure volume index (PVI)

Case 2: Model Calibration Results (1)



Case 2: Model Calibration Results (2)



Case 2: Model Fitness (MAE/MAD) by patient, type of challenge, challenges/session, length of session, mean ICP

	P004	P006	P007	P201	P202	P204	P205	P206	P207	All
MAE/ MAD	.53	.43	.99	1.06	.45	.52	.30	.53	.96	.72
N	4	5	3	4	4	1	1	1	1	24

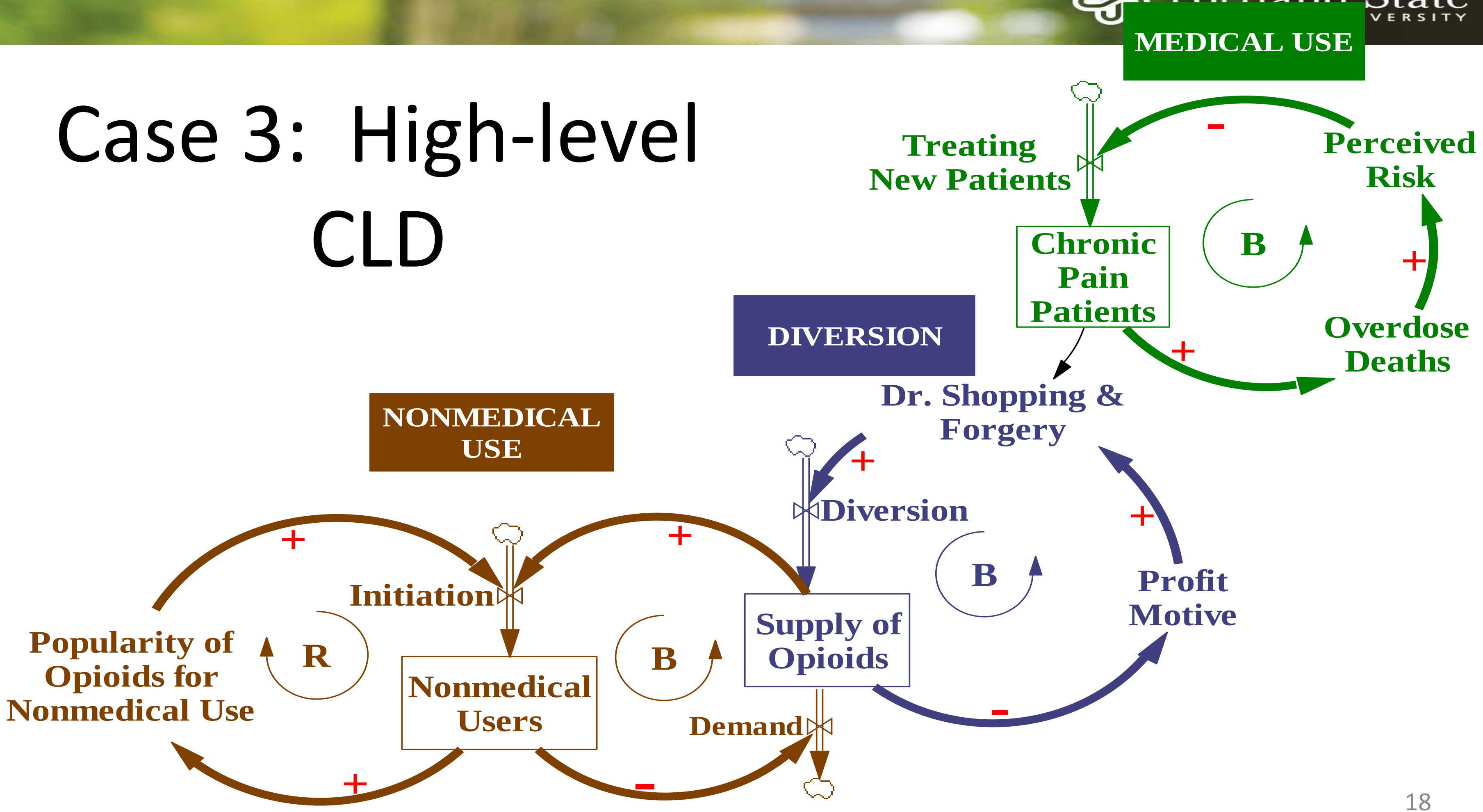
	Only HOB Challenges	Only RR Challenges	HOB and RR		<= 3 Challenges	>=4 Challenges
MAE/ MAD	.89	.50	.61		.84	.66
N	14	3	7		10	14

	Length of Session (minutes)					Mean ICP for Session (mmHg)		
	<=40	41-60	61-80	>80		Low (<12)	Medium (12-18)	High (>18)
MAE/ MAD	.54	.62	.69	.93		.47	.77	.91
N	5	9	6	4		8	10	6

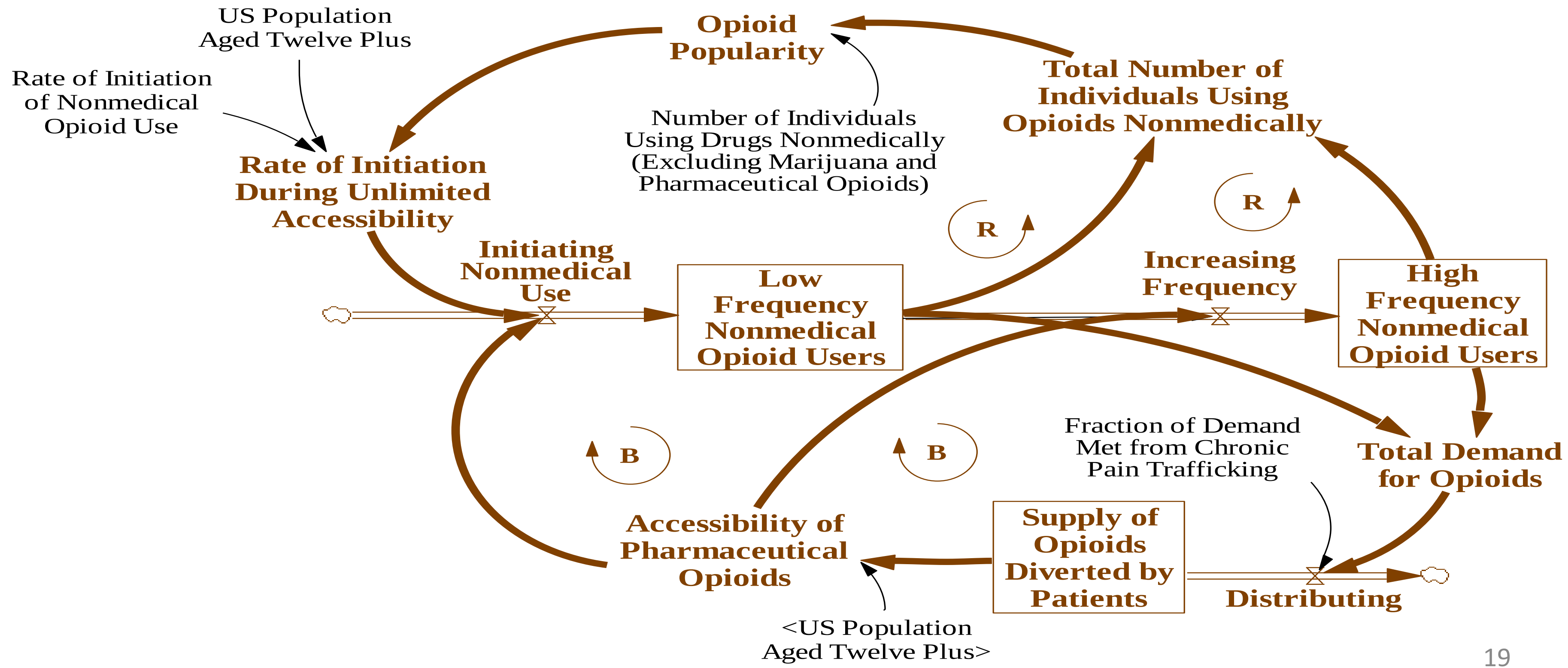
Case 3: Opioid Diversion & Abuse

- Motivation: dramatic rise in the nonmedical use of pharmaceutical opioid pain medicine and fatal overdoses; ineffective government policies and regulations
- SD models often used to study health policy
  - Homer 1993, Jones et al. 2006, Cavana and Tobias 2008, Milstein et al. 2010, among many
- Modeled medical use of pharmaceutical opioids to treat pain, drug diversion, and nonmedical use/outcomes
- 7 state variables, 90 support variables, 40 parameters
- Data from literature and other public sources
  - Direct empirical support for 12 params. indirect for 17 more
- Expert panel judgment for model structure and parameters lacking empirical support
- All but two highly influential parameters had some degree of empirical support

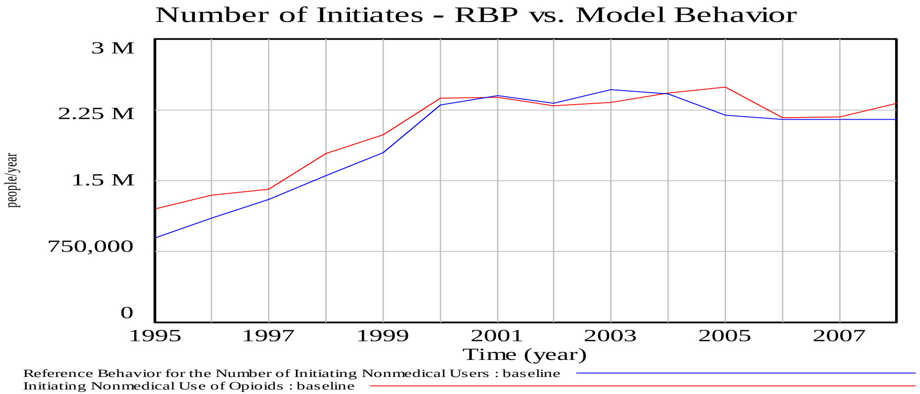
Case 3: High-level CLD

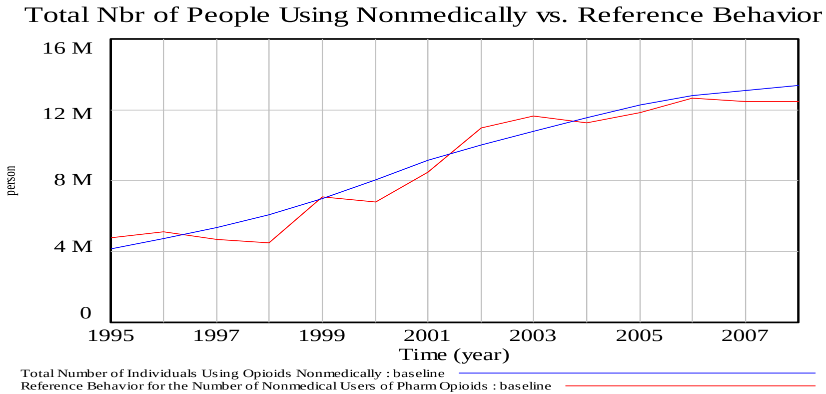


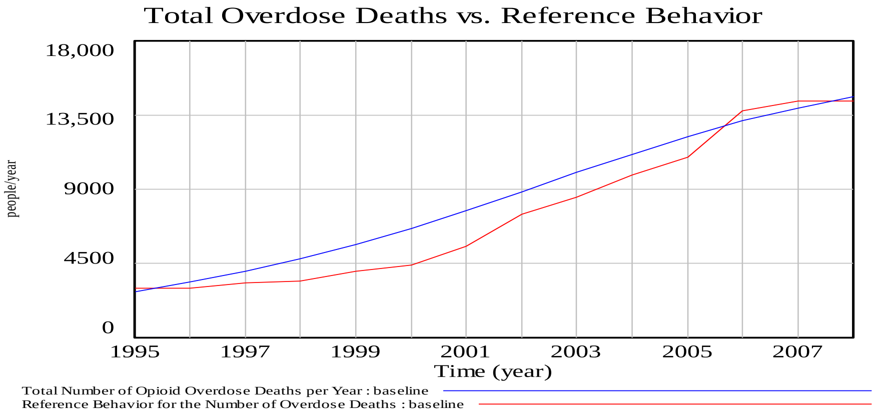
Case 3: SFD for Medical Use Sector



Case 3: Model vs. Reference Behavior



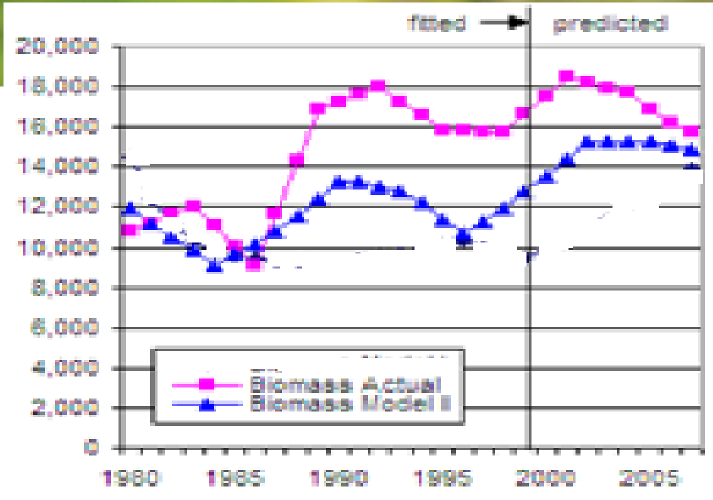


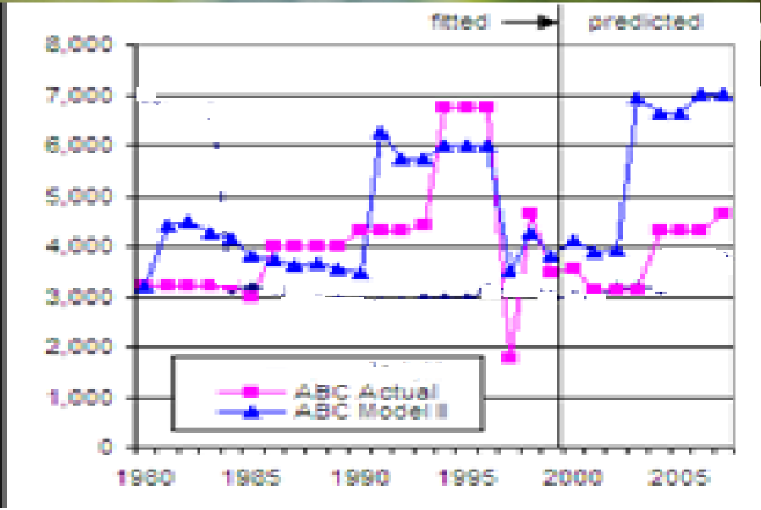


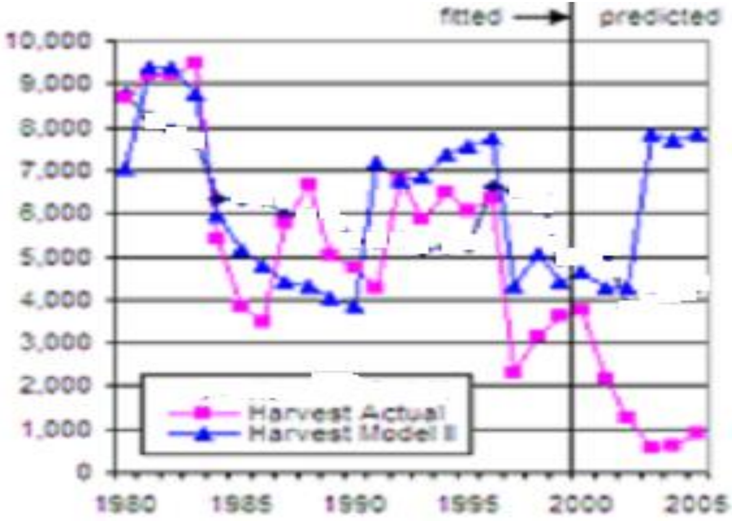
Mean Absolute Percentage Error (MAPE): 10%, 9%, 22%

Case 1: New Data

					Decision Table	
	Harvest	Spawning Biomass	ABC	MSY (OY)	Moderate Catch (F50%)	**Likely Sp. Biomass**
1992		18,000				
1995		15,822				
1998		15,735				
1999		16,955				
2000	3735	17,909	3539			
2001	2142	18,467	3146			
2002	1260	18,783	3146			
2003	551	16,324	3146			
2004	618	17686	4320			
2005	892	16915	4320		4940	**17,232**
2006				4680 (4548)	4743	**16,169**
2007					**4634**	**15,717**







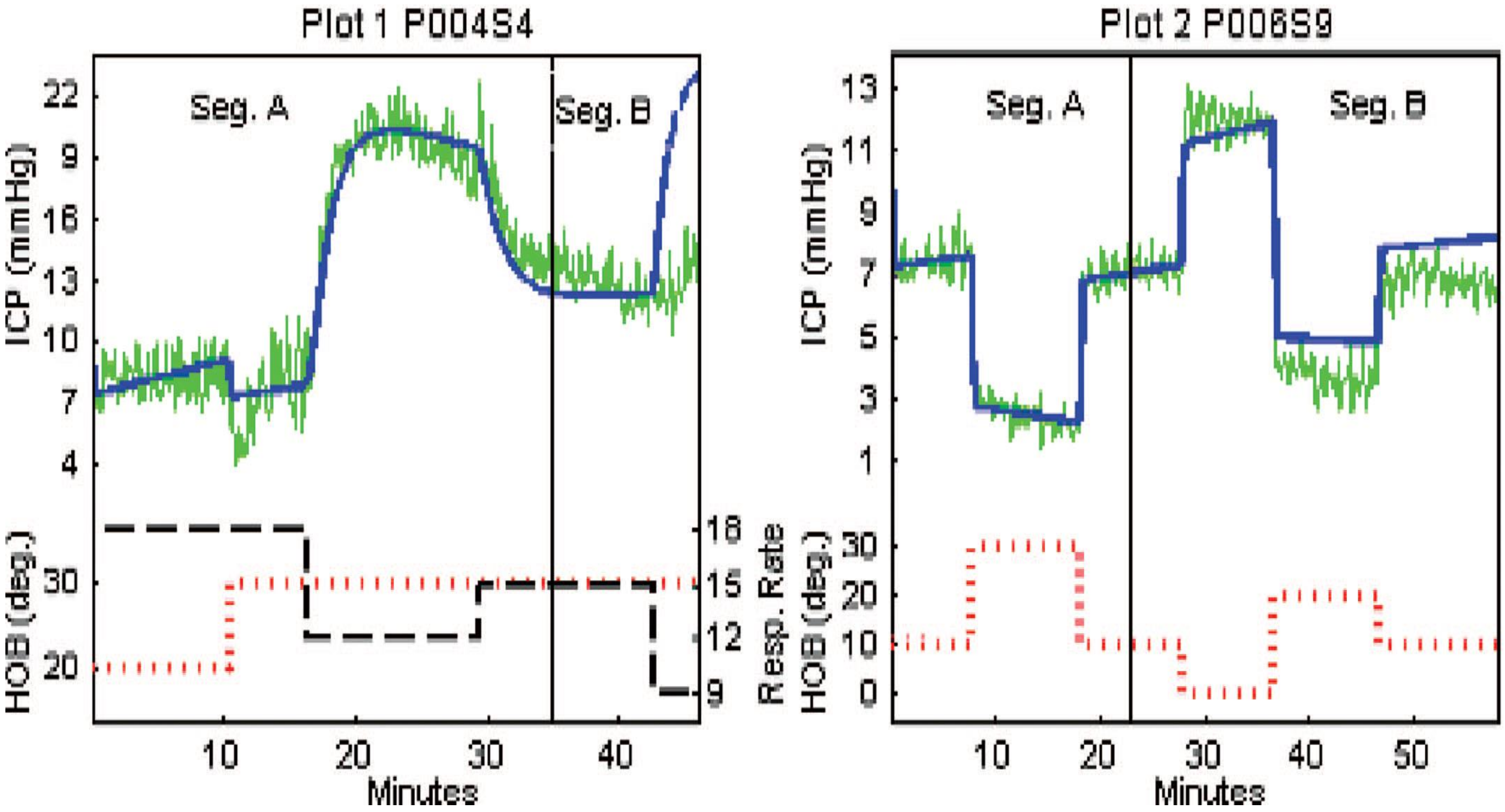
Case 1: Prediction Accuracy

	Spawning Biomass	ABC	Harvest	N
Model Fit Error	19%	24%	27%	20
Model Prediction Error	14%	51%	601%	6 for Harvest, 8 for SB and ABC

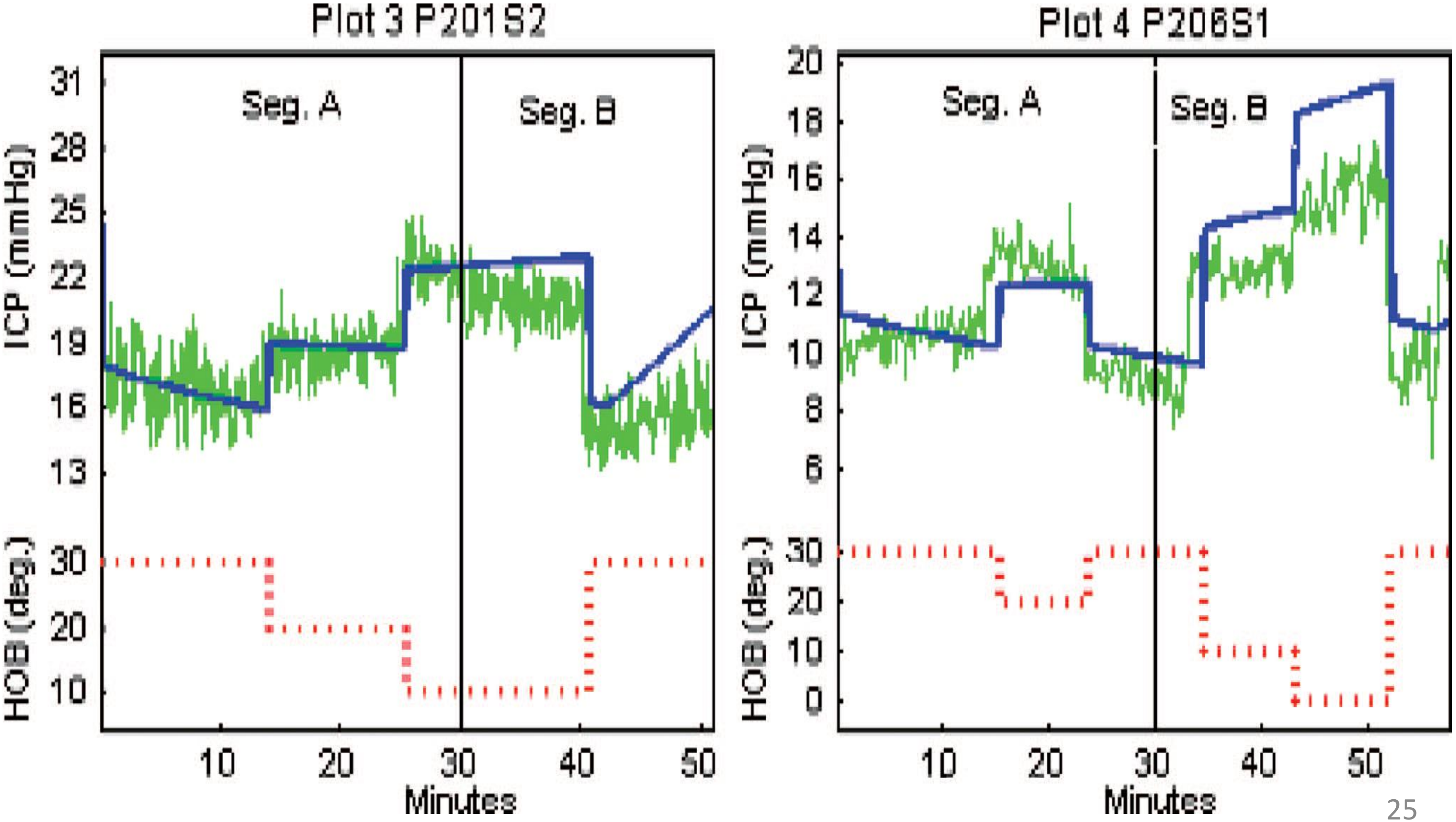
Case 1: Prediction Discussion

- Model did not capture regulatory agencies behavior
- Small changes → significant effect
  - *Spawning biomass* levels indicate “normal” fishing: *ABC* = 18% of *mature fish*
  - But, regulators chose to leave the fishery as “precautionary” w/ABC = 12%
  - This accounts for much of the model prediction error for *ABC*
- Results question whether endogenously modeling fishery regulation is possible
  - Regulators use judgment and do not set rules based only on the numbers
    - Big challenge for modelers striving to model fishery regulatory processes
  - E.g., closing a fishery because a co-mingled fishery is in danger
    - Model boundary issue
  - Supports Pilkey and Pilkey-Jarvis (2007) assertion that environmental scientists “cannot predict the future” even with (or perhaps because of) their reliance on quantitative models

Case 2: Example Prediction Results (1)



Case 2: Example Prediction Results (2)



Case 2: Prediction Error w/in Segment (MAE/MAD)

Patient	Best Fit	Predicted	N
P004	.43	1.88	3
P006	.48	.59	5
P007	.83	3.49	3
P201	1.81	1.79	4
P202	.38	3.50	2
P204	.81	2.57	2
P205	.76	1.43	1
P206	.62	1.61	1
P207	.94	1.03	1
Total	.82	1.90	22

Portland State University

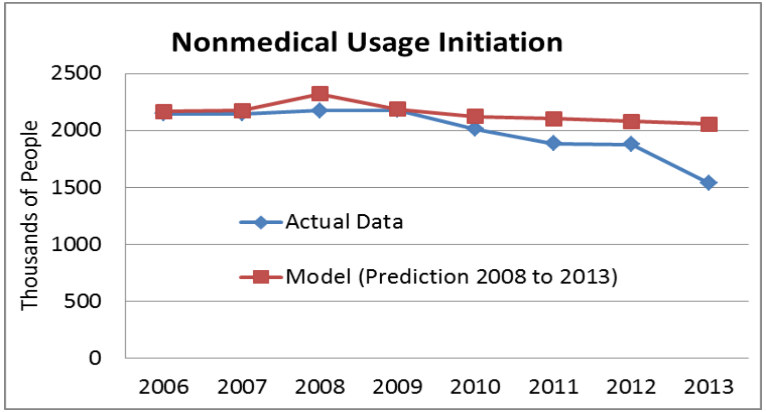
Case 2: Prediction Error between Sessions

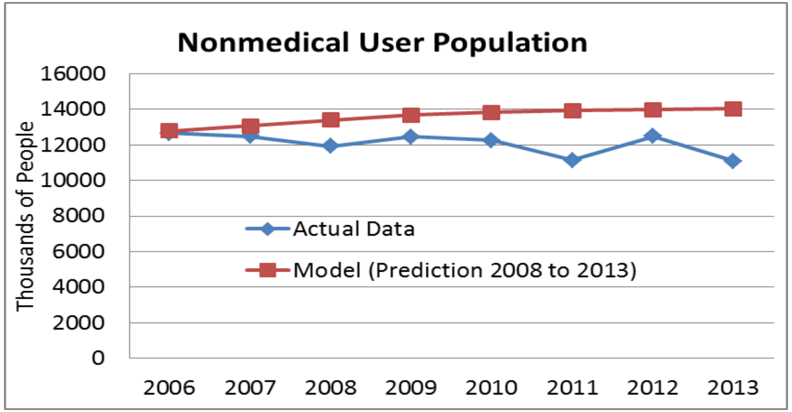
Patient	Prediction Error (MAE/MAD)	N
P004	1.93	6
P006	1.99	10
P007	2.34	3
P201	2.99	6
P202	2.88	6
Overall	2.41	31

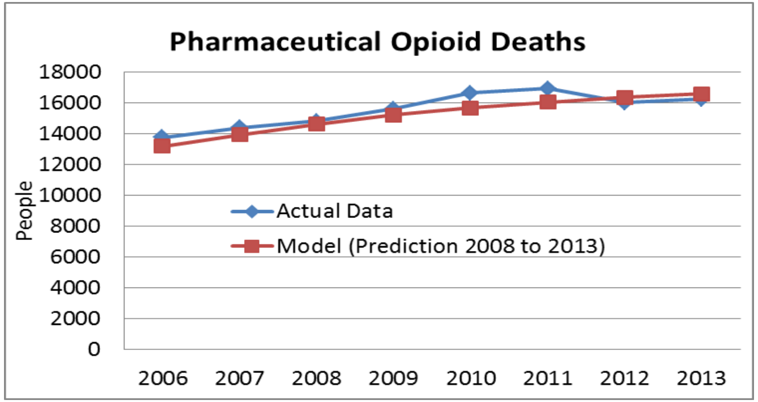
Case 2: Discussion

- Model prediction error for ICP is far too large to be clinically useful
  - Disappointing, as model fitness to RBP was much better
  - Fitness to RBP may not indicate model's utility for prescriptive analysis
- Prediction is hard, especially for human physiology
  - Due, in part, to high degree of non-stationarity
- Ultimately, the patient-specific model research was abandoned
  - Due to high intra-patient non-stationarity / variability
  - Though well-known to clinicians and easily seen in the data, it was the attempt to make predictions that forced researchers to revise their expectations…

Case 3: Prediction Errors (2009-2013)







- 5-year MAPE
  - 7%, 14%, 3%

Case 3: Discussion

- Five-year prediction errors of 7%, 14%, and 3% seem respectable
- But, these predictions did not capture the reduction in initiation and number of nonmedical users
- Might not be a bad thing altogether, because the baseline model assumed no policy change
  - Whereas, in 2011, the most abused medicine, OxyContin©, was re-issued in a truly tamper-resistant formulation, and since then, it has been less diverted and abused
  - Also, prescription drug monitoring programs are now operating in 49 states
    - Prescribers can check to see if their patients are getting medicines from other docs; and, some prescribers are being more cautious
- Making predictions and checking their accuracy added value beyond the replication of reference behavior

Study Limitations

- Was based on three projects led by a single researcher
  - Findings could be highly biased and non-representative
  - Future work should involve models created by multiple researchers to avoid potential biases and idiosyncrasies
- Method was retrospective, subjective, and did not employ a refutable hypothesis coupled with earnest efforts to refute that hypothesis
  - Such an approach could strengthen support for the assertion that prediction tests are the quintessential model tests for SD-based policy/prescriptive models

Conclusion

- When model objectives include forward-looking policy evaluation, testing prediction accuracy can be important
- When automated calibration algorithms are used, it may be sufficient to hold back part of the data, calibrate model using a training subset, and measure prediction performance using the holdout sample
- If manual calibration is used, modeler must be blind to recent outcomes, make predictions of recent outcomes, get the actual data, and measure prediction performance

A Nagging Worry

- Do complex models that more fully reflect system interconnectivity and dynamics actually predict system behavior better?
  - Conventional wisdom, and likely empirical evidence, may suggest otherwise
  - When forecasting, simple models often outperform complex models
- These modeling cases are thought-provoking, and seem to indicate that complex models should be used with considerable caution…

Further Reflections

- More complex SD models can lead to deep insights into structure and behavior that are likely not possible with simple non-parametric models
  - The point is not that SD models should be used for making predictions, but rather that prediction testing is useful to test whether a policy-oriented model is ready to be deployed
- Hmmmm. Does “policy analysis” actually *require* prediction?
  - Certainly prescriptive models (such as the ICP dynamics model) must be able to predict
  - But do policy analysis models need to make accurate predictions?
  - Could a model with poor numerical predictive ability still make useful *qualitative* predictions that lead to deep and useful insights?
    - If so, then how might a modeler assess *qualitative* predictive utility?