Prediction--the Quintessential Policy Model Validation Test

Wayne Wakeland
Portland State University, Systems Science Ph.D. Program
P.O. Box 751, Portland, OR 97207
Tel: 503-725-4975/Fax: 503-725-8489
Email: wakeland@pdx.edu

Abstract

It is essential to objectively test how well policy models predict real world behavior. The method used to support this assertion involves the review of three SD policy models emphasizing the degree to which the model was able to fit the historical outcome data and how well model-predicted outcomes matched real world outcomes as they unfolded. Findings indicate that while historical model agreement is a favorable indication of model validity, the act of making predictions without knowing the actual data, and comparing these predictions to actual data, can reveal model weaknesses that might be overlooked when all of the available data is used for model development. Although this finding is based on just three cases, the value of using prediction to validate models, as recommended by leaders in the field, is compellingly demonstrated. The implication for decision makers is to be cautious about analyses made using models that have been tested only against historical data. The primary contribution is to clearly demonstrate the oft-prescribed but less often performed validation method of testing model predictions against reality.

Keywords: model testing, fishery management, intracranial pressure, drug abuse

1. Introduction

Models must, of course, be well suited to their intended application. Consequently models used to evaluate the potential ramifications of alternative policies must be able to predict to some useful degree how the system is likely to respond to these alternative policies. An important part of the testing process for such models is to determine how well the model can accomplish this task. This means measuring not only how well the model fits and explains the past behavior, but also how well the model can project behavior into the future. This paper advocates that for models used in this fashion the discrepancy between model calculations and actual data--both historical and predicted—should be calculated and reviewed. This recommendation is fully consistent with recommendations in the SD literature that models be subjected to a wide array of model tests pertinent to their intended application.

This paper asserts that policy models should be tested to determine the degree to which the important mechanisms governing dynamic behavior have been captured. A compelling way to do this would be to predict how the system will behave in the future, over the time period of interest for policy analysis, and then compare the model prediction to actual behavior as the future unfolds.

Of course, it might be possible to fully blind oneself to the recent past, and then create/calibrate the model based only on data regarding the more distant past, and then use the model predict the recent past. This is a sound idea as long as the modeler is truly blind to the recent past. This concern regarding the modeler being blind to the metrics being predicted is lessened somewhat when an automated calibration method is used, such as employing a heuristic search algorithm to determine parameter values that maximize fitness of the model calculated results to reference behavior data (regarding the more distant past in the preceding example).

However, when models are calibrated manually by the modeler, as is often the case with SD-based policy models with multiple outcome variables, it is very easy for the modeler's choices regarding parameters values and equations to be influenced by subjective knowledge of that which is being predicted. The modeler might have glanced at a graph before being blinded, or might hear something on the news indicating how the future might be unfolding. Even if the modeler very conscientiously strives to disregard such information, they could nevertheless be subconsciously influenced.

Therefore the most compelling model prediction tests involve predicting future behavior, and then waiting until the future unfolds in order to complete the test. If the prediction period is the recent past that is currently unknown, then it could be possible to wait until after the model predictions have been made before proceeding to acquire the reference behavior data pertaining to the recent past.

Section 2 provides a brief summary of key literature on model testing and using prediction testing for model validation. The primary method used to support the paper's assertion, discussed in Section 3, is to examine three specific cases where system dynamics models were developed, calibrated against reference data, and used to make predictions. In each case, additional data was collected to in order to determine the accuracy of the predictions. The topics of the cases are fisheries management (Wakeland 2007), elevated intracranial pressure following traumatic brain injury (Wakeland 2009), and the diversion and misuse of pain medicine (Wakeland 2012, 2013). Section 4 summarizes the results, and Section 5 provides discussion, interpretation, and conclusions.

2. Background

Model testing was historically referred to as model verification and validation, but more recently authors have shifted emphasis away from validation in order to avoid the possible impression that a model can be declared valid or invalid by running a set of tests. Instead we say that a model has been well tested or that we have established the applicable domain for which the model performs well. Within in SD field model testing has received considerable attention from its inception and recently (cf Barlas 1996, Coyle and Exelby 2000, Sterman 2000, Saysel and Barlas 2006, Groesser and Schwaninger 2012). The concept of testing a model's predictive capability is discussed in some detail, but few examples have been provided.

3. Method

For each of the three cases examined, the context is summarized, including key aspects of the model, the model calibration approach, and information regarding how well the model calculated results fit the reference behavior.

3.1 Fishery Regulation

A fisheries management case study is appropriate because fishery regulatory agencies have found that stopping the decline of fish populations is very challenging, and therefore many models have been built to address this challenge. Populations of rockfish, for example, dropped dramatically in recent decades; and, since 1983, rockfish landings have decreased 78% and catch limits for various species of rockfish have been reduced by 78%-89%. In 2000, the West Coast ground fish fisheries were declared a federal disaster (Ecoworld 2000). The decline in fish stocks is considered by many to be the consequence of ineffective natural resource management and short-term policies that resulted in a larger fishing fleet than could be supported long term.

Mathematics, statistics, computer simulation, and System Dynamics (SD) have all been used to model fishery management systems. Schaefer (1954) provided the classic dynamic (differential equation) model for fish biomass as a function of pristine (unfished) biomass, intrinsic rate of increase, fishing effectiveness, and fishing effort. Applications of SD to fisheries management are plentiful (c.f., 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, and Wakeland 2007).

3.1.1. Fishery Regulation Case Model

This case describes an SD model of the Pacific yellowtail rockfish developed in 2003 based on data up through 2000. The model included fish populations, fishery regulation, and fishing activity, including the degree of compliance with regulations. Many parameter values for the model were taken from the literature and from reports published by the Pacific Fisheries Management Council (Pcouncil 2003). Other parameters were estimated by calibrating the model to best fit to a portion of the reference data. Findings from the research supported the generally accepted rules of thumb regarding maximum sustainable yield, suggested that more frequent updates to acceptable biological catch based on more frequent stock assessment studies would help to stabilize the fishery, and noted that shortening management response time for adjusting fleet capacity would also be highly beneficial. and then updated in early 2007 using this same data. Figure 1 shows a high level causal loop diagram for the model.

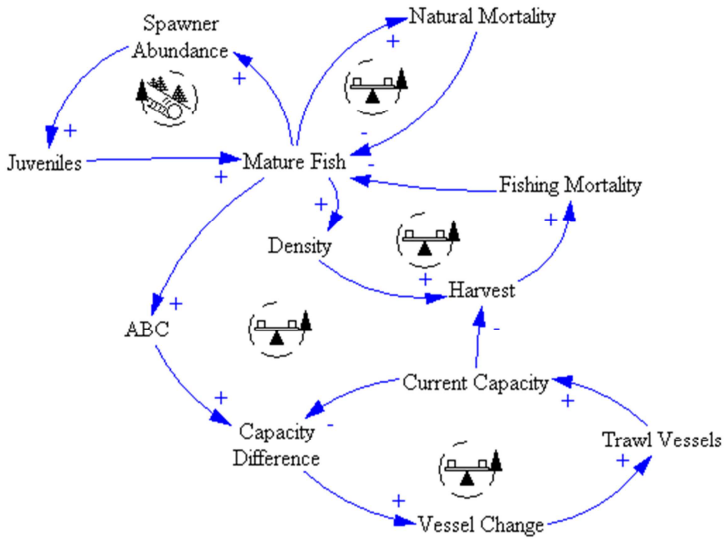

Figure 1. High level cause loop diagram for the fisheries case study model

The model was initialized in steady state circa 1980, and tested to determine parameter sensitivity and the domain of applicability of the model, which is established by finding the extreme values at which the model ceases to function properly. The model was then run for a 20-year period in order to endogenously calculate values for the outcome variables from 1980 to 2000. Six model parameters were adjusted experimentally to achieve the best fit to the historical data for this period. Table 1 lists the parameters that were adjusted, their plausible range, and their final values.

Table 1: Fishery case model parameters adjusted to achieve best fit with historical Metrics

Parameter	Plausible Range	Final Value
Surviving into juveniles per spawner w healthy ocean (#)	1 - 5	3.5
Recruit base annual mortality fraction (#)	.1 - .3	.23
Initial value for Mature Fish (#)	20 – 30M	27M
Pre '85 enforcement fraction (#)	.5 - .8	.7
Fishers Participation Change Response Time (Yrs.)	2 – 5	3
trip limit effectiveness divisor (fish/vessel)	200 – 300K	250K

Model versus actual data was plotted for the three key metrics: spawning biomass (*Biomass*), acceptable biological catch (*ABC*), and *Harvest*; and the mean absolute percentage error (MAPE) was calculated for each metric. Figure 2 shows model results vs. actual data for *Biomass*, *ABC,* and *Harvest*. MAPE was 35% for *Biomass*, 24% for *ABC*, and 27% for *Harvest*.

Predicted values for the key metrics for the period from 2001 to 2005 were saved for use in subsequent analyses (Section 4.1.1).

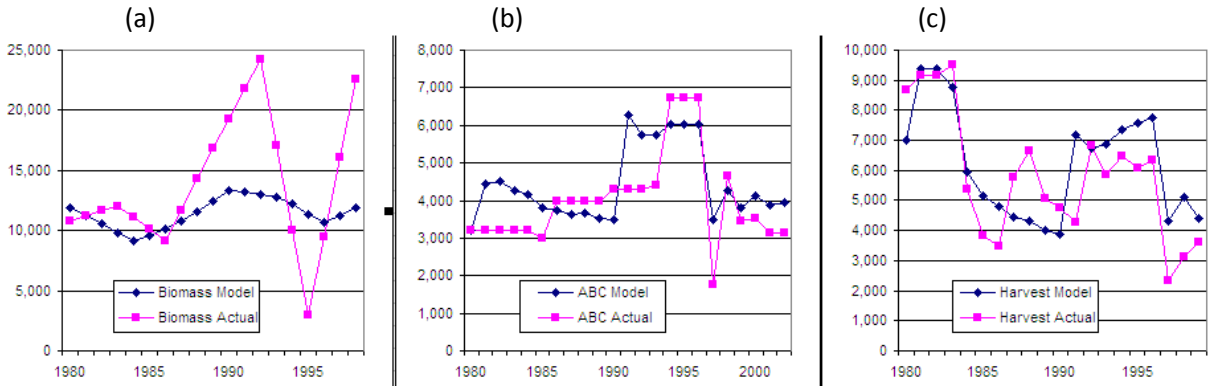

Figure 2. Model fitness for key metrics: Biomass (a), ABC (b), and Harvest (c)

3.2. Intracranial Pressure (ICP) Case

This case is relevant because outcomes for elevated ICP following traumatic brain injury (TBI) remain mixed despite many scientific and clinical advances, and TBI remains leading cause of death and disability in children, with the death rate for severe TBI ranging from 30-45% at major children's hospitals (White 2002). Although many sophisticated computer models have been created (see Wakeland 2008 for a review), much of the necessary data remain difficult to obtain. Researchers have estimated parameters by calibrating them to fit patient-specific clinical data (cf., Ursino and Lodi 1997, Ursino and Magosso 2001, Wakeland et al. 2005, Hu et al. 2007), and, in some cases, excellent results have been reported (Ursino, Lodi, Russo 2000 and Ursino, Minassian, Lodi et al. 2000). However, prior to Wakeland et al (2009) no study had reported the capability of these models to make predictions, although sometimes model calculations have been referred to as predictions when the aim was to match ("predict") reference data (e.g., Ursino, Minassian Lodi et al. 2000).

A system dynamics model of intracranial pressure dynamics was developed during the period from 2003 to 2006 and calibrated to specific patients based on prospective data that were collected from pediatric patients being treated for traumatic brain injury. Under an IRB-approved protocol, the patients were given mild challenges (by changing the head of their bed from zero to 30 degrees and vice versa, and changing their respiration rate to create mild hyper-ventilation and mild hypo-ventilation.), and their physiological responses carefully measured and recorded (Wakeland et al. 2005). The objective was to determine if patient-specific models could be created that could predict the patient ICP response to interventions. Such models could be used to evaluate potential treatments in-silico prior to being administered to patients.

Data was collected on nine TBI patients and included 24 testing sessions in total. Data from early in a single long session or from prior sessions was used to estimate patient-specific parameter values. This was done using an algorithm (see Figure 5) that minimized the squared error between the model-simulated ICP values for a specific "challenge" session and the actual ICP during the session. Another approach could be to use Kalman filters as reported by Hu et al. (2007). The resulting patient-specific models were used to predict patient's ICP response to similar but different interventions at other points in time, either later in the same session or during subsequent sessions.

3.2.1 ICP Dynamic Model

Figure 3 shows the primary stocks and flows in the ICP dynamic model. The model was developed using an SD modeling package, subjected to a fully battery of sensitivity tests (Wakeland and Hoarfrost 2006), and then implemented in Matlab (Figure 4) in order to estimate the patient specific parameters using the process shown in Figure 5. The algorithm adjusted the following ten parameters in order to achieve best fit between the model-calculated ICP and the ICP data collected for the patient during a particular challenge:

- 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)

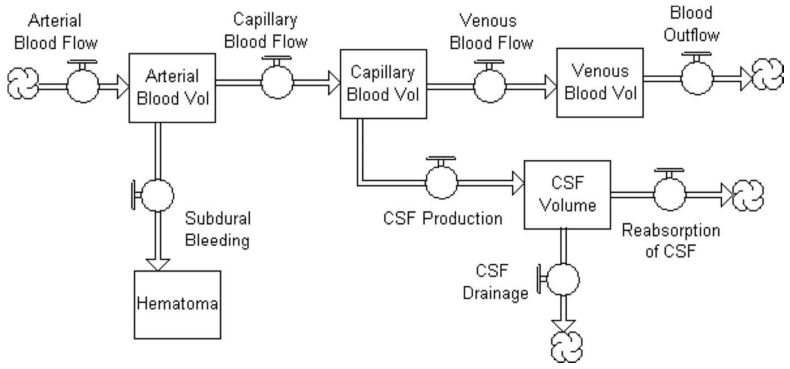

Figure 3. Primary stocks and flows in the ICP dynamic model

Figure 6 shows the actual ICP data that were collected for the 24 challenge episodes with the patient-specific ICP model-calculated ICP superimposed. In some cases model fitness was very good and in other cases not good at all. Overall, the average mean absolute error (MAE) in the model-calculated ICP was 1.9 mmHg compared to an avg. mean absolute deviation (MAD) of 3.1 mmHg. Thus, the overall MAE/MAD is .61, which is not great, but could be potentially clinically useful.

Table 2 gives additional details regarding the fit (MAE/MAD), for each patient, by type of challenge, by the number of challenges given during a session, by the length of the session, and by the mean ICP during the session. These data show that model fitness is highly variable by patient, better for respiration rate challenges, better for short sessions, and better for sessions in which ICP is only mildly elevated. Overall, these results were considered to be encouraging.

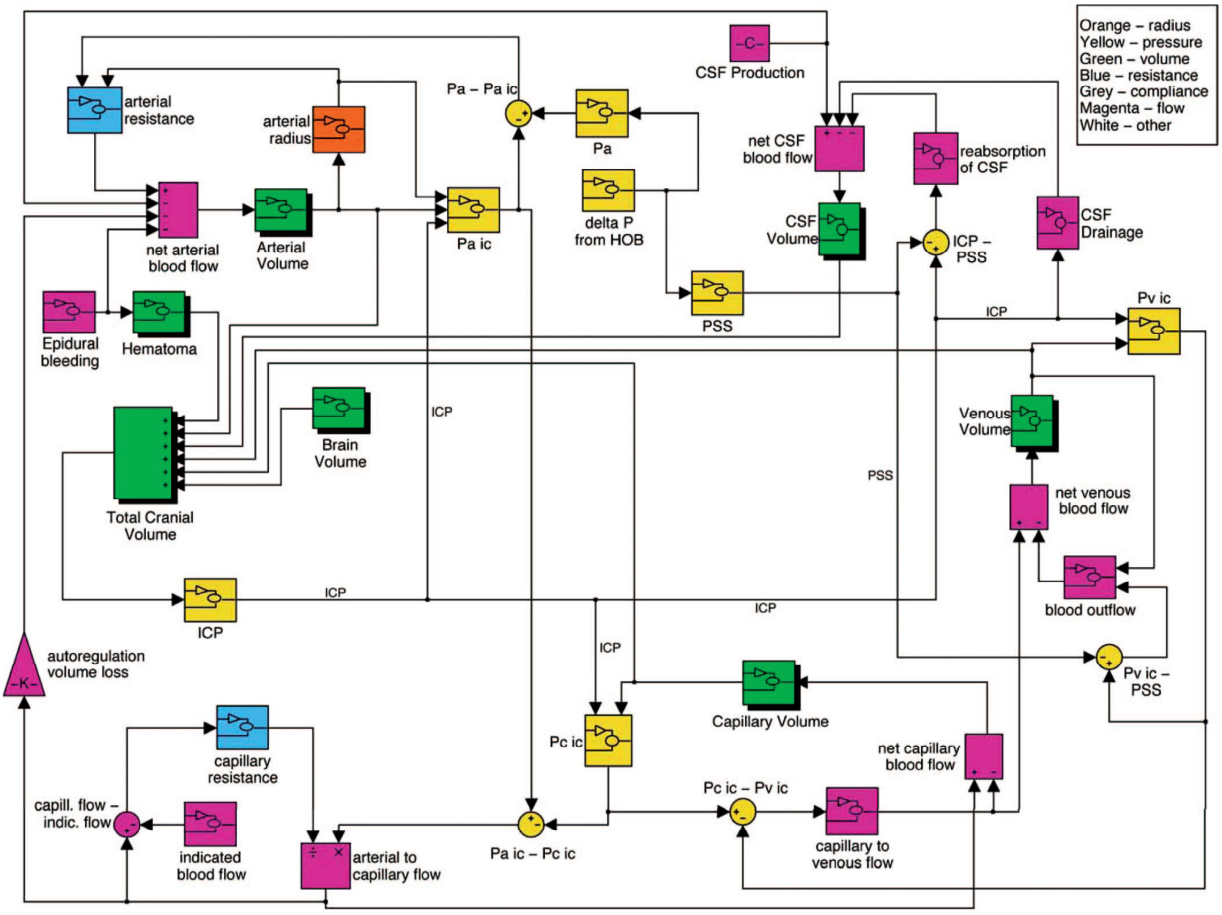

Figure 4. Simulink diagram of ICP dynamic model used for parameter estimation and prediction

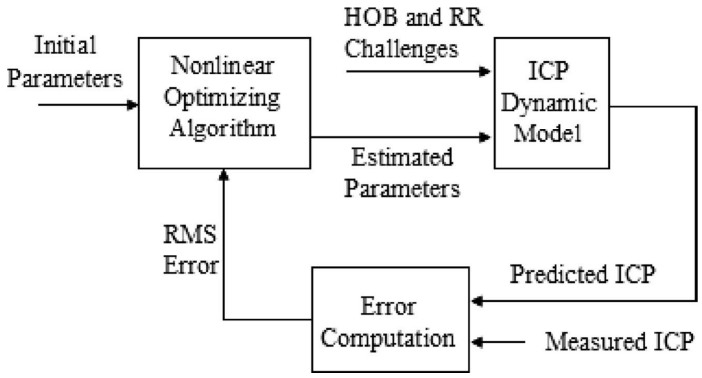

Figure 5. Parameter estimation process for create patient-specific ICP dynamic models

[moved to supplement due to size constraints]

Figure 6. Observed and modeled intracranial pressure (*ICP*) using individually fit parameter values. The observed ICP waveform is the green jagged trace. The modeled ICP is the nonjagged blue trace. The traces at the bottom show changes in the head-of-bed (dotted) and ventilation rate (dashed)

Table 2. Model fitness errors (MAE/MAD) grouped by patient, by type of challenge, by the number of challenges in a session, and by the mean ICP for the session.

	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

3.3. Opioid Diversion and Abuse Policy Model Case

The objective of the study considered in this third case was to develop a system dynamics model of the medical use of pharmaceutical opioids to treat pain, and the associated diversion and nonmedical use of these drugs. Motivation stemmed from a dramatic rise in the nonmedical use of pharmaceutical opioid pain medicine (Compton and Volkow 2006, Warner et al. 2011). Despite the increasing prevalence of negative outcomes, such as non-fatal and fatal overdoses, nonmedical use of pharmaceutical opioids remains largely unabated by government policies and regulations (Fishman et al. 2004).

SD models have frequently been applied to study health policy issues (c.f., Homer 1993, Jones et al. 2006, Cavana and Tobias 2008, Milstein et al. 2010). Figure 7 provides a sense of the structure of the SD model employed in third case. The model has seven state variables, 90 support variables including outcome metrics and policy variables, and 40 parameters. The research relied on secondary data obtained from the literature (c.f., Degenhardt et al. 2004, Fisher et al. 2004, Manchikanti et al. 2006, Warner et al. 2009) and from other public sources, such as the National Survey on Drug Use and Health and the CDC, for the period 1995 to 2008 (c.f., Colliver et al. 2006, SAMHSA 2007, Governale 2008). Twelve of the parameters have direct empirical support, and indirect support was identified in the literature for another seventeen. An expert panel provided recommendations regarding the model structure and the remaining model parameters. The model was subjected to a full battery of tests, with

sensitivity testing being particularly informative. All but two of the highly influential parameters had at least some degree of empirical support.

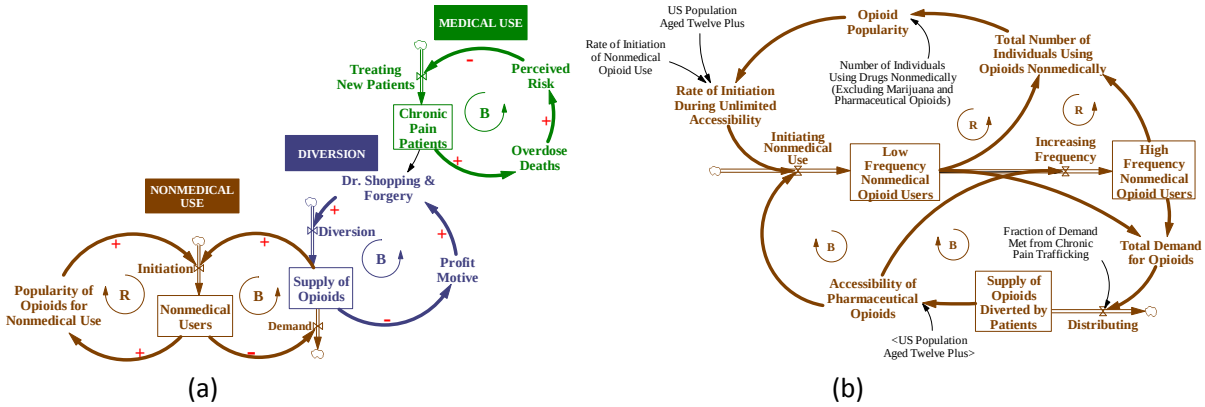

Figure 7. Diagrams illustrating key aspects of the opioid policy model. (a) A high level diagram showing the three model sectors and their interconnections. (b) Details the Nonmedical Use Sector to illustrate the approach

The model was calibrated experimentally by adjusting parameter values within their plausible range. Parameters with solid empirical support were not changed. Figure 8 shows the degree of fitness of the model to the reference data for the annual number of persons who initiated nonmedical use of pharmaceutical opioids, the total population on nonmedical opioid users, and opioid-related overdose deaths. Model fitness was acceptable for the user populations, with errors of 9% and 10%, but was less satisfying for overdose deaths, with 22% error.

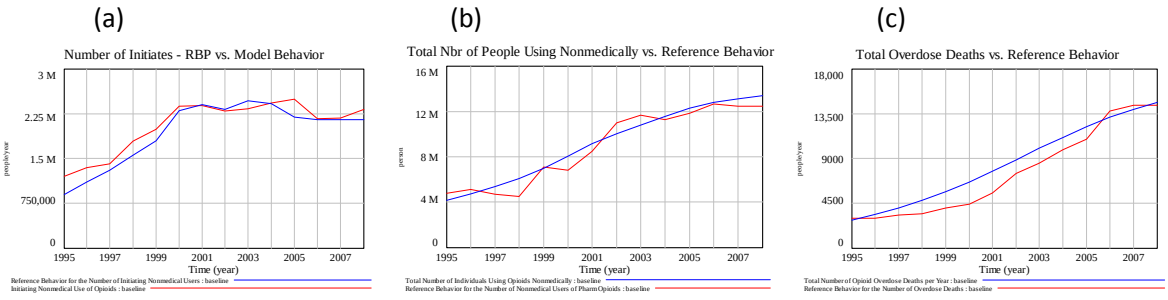

Figure 8: Opioid model fitness to reference behavior data. (a) Number of individuals initiating nonmedical opioid use per year (Mean Absolute Percentage Error [MAPE] 10%). (b) total nonmedical users of prescription opioids (MAPE 9%), (c) total prescription opioid overdose deaths per year (MAPE 22%).

The model was then run out to 2015 to serve as the baseline for policy analysis. The baseline forecast will be evaluated in Section 4.3 to determine the accuracy of the model's predictions. In the primary study, the impact of simulated interventions were tested and compared.

4. Results

For each case study, the baseline predictions made by the model are presented next, along with the approach used to obtain the additional data needed to assess prediction accuracy, and then each model's prediction accuracy is determined and explained.

4.1. Results for Fisheries Model Case

Data was acquired in mid-2007 regarding the predicted five-year period. These data were from two sources, the Pacific Fisheries Management Council website (Pcouncil 2007) and its Final Environmental Impact Statement (FEIS 2007), plus the Status of the Yellowtail Rockfish in 2004 (Yellowtail 2005). Table 3 provides the revised data for key variables that were gleaned from these recently released documents. Also of interest in the recently released documents was the statement that since 2003, commercial fishing for yellowtail rockfish has been substantially curtailed because this fishery co-occurs with other fisheries that are classified as depleted: the canary rockfish and widow rockfish (FEIS 2007, pg. 259).

Table 3: New Data from 2005 and 2006 Reports (Metric Tons)

					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

Table 3 also provides revised estimates for actual spawning biomass for the years 1992, 1995, and 1998. The original data points for these years were suspect due to their very high degree of variability (see Figure 2a). The revised data are much more plausible (see Figure 9a). Finally, Table 3 also shows MSY and OY (maximum sustainable yield and optimum yield) for 2006, and reasonable allowable catch estimates for 2005 to 2007. These are provided for comparison to model predictions since harvest had been suspended.

4.1.1. Model Prediction Error

Figure 9 shows the actual *biomass*, *ABC*, and *harvest* up to the present (2007), along with the fitted and predicted values from the model.

Table 4 compares the average model fit error and model prediction error for each key metric. The MAPE for *Biomass* fit error was revised to 19% instead of 35% based on the revised actual data shown in

Table 6. Prediction error for *spawning biomass* was actually less than its model fit error, which is not typically the case.

Given the fact that the fishery was closed to fishing for reasons totally external to the model, the model could not have predicted the *harvest* accurately. Nevertheless, the prediction test revealed model weaknesses that had previously not been detected and showed compellingly that model modifications

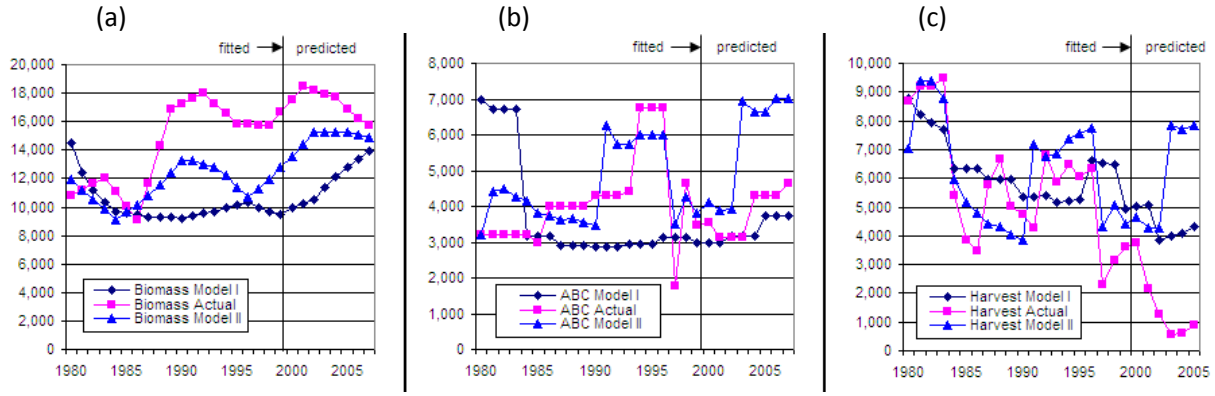

Figure 9. Predicted and actual *Biomass*, *ABC*, and *Harvest*. Please focus on the "Model II" traces and ignore the Model I traces.

Table 4: Model Fit and Model Prediction Mean Absolute Percentage Errors

	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

would be necessary before the model could confidently be used to evaluate policy scenarios. In particular, the model scope would need to be changed to include other fish species caught with the same type of gear, since regulators take this into account.

4.2. Results for ICP Dynamic Model Case

To measure the model's prediction capability, for sessions in which multiple challenges were given, the patient's response to initial challenges within the session were used to estimate parameter values, and the resulting model was then used to predict the patient response to the later challenges. Figure 10 shows four selected sample results from this process.

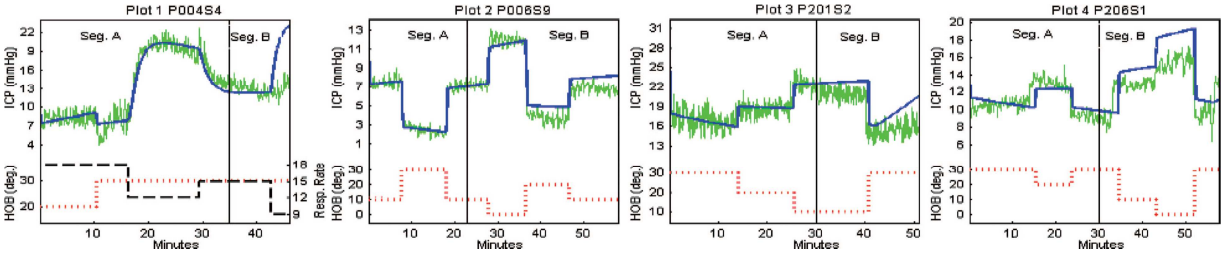

Figure 10. Each plot shows the ability of the model to predict intracranial pressure (*ICP*) within a selected session. Seg. A was used to estimate parameter values for the model, and the ICP was calculated for both Seg. A and Seg. B using these parameters (*blue, solid line*). The green, jagged line is the observed ICP. A vertical bar separates the two segments. The dotted and dashed lines show the changes in head-of-bed and respiration rate, respectively.

Table 5 summarizes the overall results of the within-session prediction tests. The average mean absolute prediction error (MAE) was a disappointing 4.0 mmHg, and the average MAE/MAD ratio was an even more disappointing 1.9.

Table 5. Within-session prediction test results by patient. Best fit is the MAE/MAD for the training sub-segment and predicted is the MAE/MAD for the predicted sub-segment. N is the number of sessions.

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

Next, the prediction error between sessions was determined by using the parameters estimated in a prior session to make predictions for how the patient would respond in subsequent sessions (see Table 6). Overall, the MAE was 6.7 mmHg, which is too poor be even remotely clinically useful, with the average MAE/MAD being 2.41.

Table 6. Prediction error between sessions. N is the number of predictions made.

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

4.3. Results for Opioid Policy Model Case

Since the original baseline predictions were made in 2011, reference data has become available for 2009-2013. Figure 11 shows plots of the model prediction compared to the actual data for the three key outcome variables. The errors in the population variables are roughly comparable to the model fitness

error for these variables (10%/7% for Initiation, and 9%/14% for User Population). The prediction error for deaths was remarkably small compared to the associated model fitness error (22%/3%).

5. Discussion

For each case, reasons are considered that might explain why the predictions were accurate or inaccurate, and what these findings imply regarding the utility of the model, next steps that might be

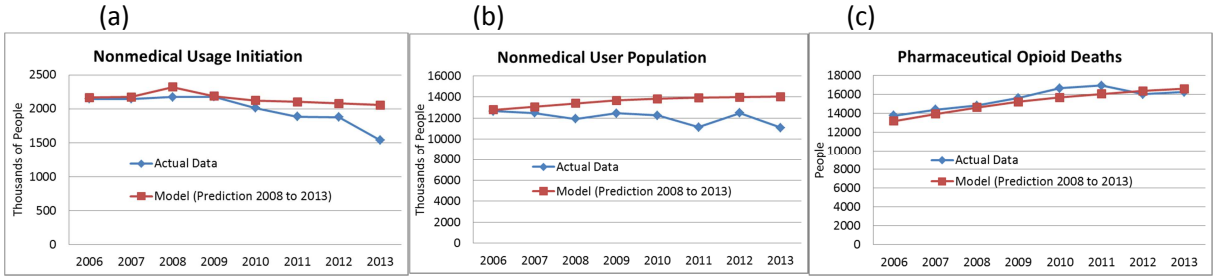

Figure 11. Prediction Error for Opioid Policy Model. (a) Nonmedical opioid usage initiation, mean absolute percentage error [MAPE] 7%. (b) Nonmedical User population MAPE 14%. (c) Pharmaceutical opioid deaths MAPE 3%.

indicated, and the key learnings. A synthesis of the findings across all three cases is provided, limitations of the study are discussed, and overall conclusions are summarized.

5.1. Fisheries Case Discussion

One possible explanation for poor predictive performance for the fisheries model with regard to *ABC*, despite have predicted *spawning biomass* reasonably accurately, is that the model did not capture the logic used by regulatory agencies. Small changes can have a significant effect on the numbers. For example, with the current levels of *spawning biomass*, regulatory practice would allow "normal fishing," meaning that *ABC* would be 18% of *mature fish*. However, if the regulators chose to leave the fishery in the "precautionary" category, ABC would be 12% of the *mature fish.* This difference would be sufficient to explain the model prediction error for *ABC*.

As is nearly always the case when using the System Dynamics method, the process of creating and working with the models was at least as useful and informative as the actual numerical results. It became clear to the researcher while working with the model that the yellowtail rockfish fishery is recovering nicely from the over fishing that took place during the early 1980's. The researcher also gained a heightened appreciation for the delicate balance that exists between the fish, the fishers, and the regulatory process. Several parameters are critical to maintaining this balance, including those shown in Table 1, especially the *Fishers Participation Change Response Time*.

This case study raises doubts regarding the prospects for endogenously modeling fishery regulation. One challenge is the presence of exogenous events that impinge on the regulatory process, such as closing a given fishery not because it is in danger, but rather because other fisheries that are co-mingled with it are in danger. Another challenge is the fact that regulators use judgment when applying regulatory rules, and do not (and should not) set rules based only on the numbers. This is a significant

challenge for those who seek to endogenously model the regulatory process, a finding that seems to be disconcertingly supportive of the recent claim by Pilkey and Pilkey-Jarvis (2007) that environmental scientists "cannot predict the future" even with (or perhaps more accurately, because of) their reliance on quantitative models.

5.2. ICP Case Discussion

Model prediction error is far too large to be clinically useful. This is disappointing, especially since the error in model fit to the historical data was much smaller. Obviously, caution warranted, and evidence is mounting that a good fit between model calculated figures and historical data may not indicate that a model will be able to make useful predictions. Certainly, prediction is hard, especially with respect to human physiology, which is notoriously non-stationary. A given patient might respond one way to a given test on day one and exactly the opposite to the same test at a later time. This was true for the prospective data collected to support the research project.

Ultimately, the research to develop patient-specific models to inform treatment was not pursued due to the high degree of intra-patient variability. Certainly this variability is well-known to clinicians, and can be seen directly in the data, but it was the attempt to make predictions that forced the researchers to recognize the problem and to revise their expectations.

5.3. Opioid Case Discussion

While the five year prediction errors of 7%, 14%, and 3% seem quite respectable, a look at Figure 11 shows that the prediction captured neither the reduction in the initiation rate over time nor the reduction in the size of the nonmedical user population over time. Actually this might not be a bad thing. The baseline model assumed no change in policies, whereas, in fact, in 2011 the most popular pharmaceutical medicine for abuse, OxyContin© was re-issued in 2011 with an improved tamper-resistant formulation, and all indications are that this new formulation is much harder to abuse and as a result this medicine is no longer abused as it was in the past. Additionally, prescription drug monitoring programs (Fishman et al. 2004) are now in place in all but one state, and these programs, which make it possible for prescribers and pharmacies to check to see if patients might be receiving medicines from multiple prescribers, are benefited from technology, and preliminary evidence indicates that prescribers are being more cautious.

But it seems clear that for this particular case, making predictions and checking to see whether or not the predictions are accurate was informative in ways that go well beyond the benefits of striving primarily to replicate reference behavior.

5.4. Study Limitations

This study is based on three modeling projects that were initiated and led by a single researcher. It is entirely possible that the findings are highly biased and non-representative. Future work should involve a representative sample of researchers and make every effort to avoid biases and idiosyncrasies. The method employed was subjective and did not involve the establishment of a refutable hypothesis coupled with earnest effort to refute that hypothesis. It would be useful to use such an approach to test the assertion that prediction tests represent the quintessential model test for SD-based policy models.

5.5. Conclusion

Overall, these results underscore the importance of measuring how well models actually fit the reference data during the model testing phase. Furthermore, in situations where model objectives include forward-looking policy evaluation, these tests should be extended to include determination of prediction accuracy. In some cases, where automated calibration algorithms are used, it would be sufficient to hold back part of the data and then to calibrate the model using a training subset of the data, and then measuring how well model predictions fit the outcome data in the holdout sample that was not used to inform the model calibration process. In other cases, one could deliberately remain blind to recent outcome data, make predictions, then acquire the recent data, and compare model predictions to the new data.

One would like to believe that a more complex model that reflects the complex web of interconnections in the system could better capture the complex dynamics and therefore be able to more accurately predict future behavior. Conventional wisdom, and likely empirical evidence, may suggest otherwise. However, for forecasting purposes, simple models often do outperform complex models. The three cases shown here could be construed as raising doubts about the utility of complex models. At the very least the results are thought-provoking.

On the other hand, as these cases also demonstrate, complex SD models are capable of generating deep insights into structure and behavior that are likely not possible with simple non-parametric models. The point of this paper is not to say that SD models should be used for making predictions, but rather that prediction testing is an ideal way to test policy analysis-oriented models to determine whether or not they are ready to be employed for their intended purpose.

Interesting questions come to mind. Does “policy analysis” require prediction? Certainly prescriptive models (such as the ICP dynamics model) must be able to predict. But do policy analysis models need to be able to make accurate predictions? Could a model with poor numerical predictive ability still be capable of making useful *qualitative* predictions that lead to deep and useful insights? If so, how does a modeler go about assessing the qualitative predictive capability of a dynamic model?

REFERENCES

Barlas, Y (1996). Formal Aspects of Model Validity and Validation in System Dynamics. *System Dynamics Review 13*(3):183-210.

Brekke A and E Moxnes (2003). Do numerical simulation and optimization results improve management? Experimental evidence. *J. of Econ. Beh. & Org*. *Vol 50*: 117-131.

Cavana R and M Tobias (2008). Integrative system dynamics: Analysis of policy options for tobacco control in New Zealand. *Systems Research and Behavioral Science*, *25*, 675-694.

Colliver J, L Kroutil, L Dai, J Gfroerer (2006). *Misuse of prescription drugs: Data from the 2002, 2003, and 2004 National Surveys on Drug Use and Health* (DHHS Publication No. SMA 06-4192; Analytic Series A-28). Rockville, MD: Substance Abuse and Mental Health Services Administration, Office of Applied Studies.

Compton W and N Volkow (2006). Major increases in opioid analgesic abuse in the United States: Concerns and strategies. *Drug and Alcohol Dependence*, *81*, 103-107.

Coyle, G and D Exelby (2000). The validation of commercial system dynamics models. *System Dynamics Review 16* (1):27-41.

Degenhardt L, W Hall, M Warner-Smith, M Lynskey (2004). Illicit drug use. In M. Ezzati, A. D. Lopez, A. Rodgers & C. J. L. Murray (Eds.), *Comparative quantification of health risks: Global and regional burden of disease attributable to selected major risk factors* (Vol. 1, pp. 1109-1175). Geneva, Switzerland: World Health Organization.

Dudley, R and C Soderquist (1999). A Simple Example of How System Dynamics Modeling Can Clarify and Improve Discussion and Modification of Model Structure. Written version of presentation to the 129th Annual Meeting of the American Fisheries Society, Charlotte, North Carolina.

Dudley, R (2003). A Basis for Understanding Fishery Management Complexities, Proc. 21st Int'l Sys Dyn Conf.

ECOWORLD (2000) Ecoworld Global Environmental Community, *West Coast Groundfish Fisheries declared 'Federal Disaster*. (http://www.ecoworld.org/Home/articles2.cfm?TID=84)

FEIS (2007). http://www.pcouncil.org/groundfish/current-season-management/past-management-cycles/2007-2008-final-environmental-impact-statement/ (accessed 3/17/15)

Fischer B, S Brissette, S Brochu, J Bruneau, N el-Guebaly, L Noël, D Baliunas (2004). Determinants of overdose incidents among illicit opioid users in 5 Canadian cities. *Canadian Medical Association Journal*, *171*, 235-239.

Fishman S, J Papazian, S Gonzalez, P Riches, A Gilson (2004). Regulating opioid prescribing through prescription monitoring programs: Balancing drug diversion and treatment of pain. *Pain Medicine*, *5*, 309-324.

Ford A (1999). *Modeling the Environment: An Introduction to System Dynamics Modeling of Environmental Systems*. Ch. 14. Island Press.

Governale L (2008b). *Outpatient drug utilization trends for oxycodone products* [PowerPoint presentation]. Retrieved from http://www.fda.gov/ohrms/dockets/ac/08/slides/2008-4395s1-04-FDA-Governale.ppt

Groesser, S and M Schwaninger (2012). Contributions to model validation: hierarchy, process, and cessation. *System Dynamics Review 28*(2): 157-181.

Holland, D and R Brazee (1996). Marine Reserves for Fisheries Management, *Marine Resource Economics*. [need vol, pgs]

Homer J (1993). Projecting the impact of law enforcement on cocaine prevalence: A system dynamics approach. *Journal of Drug Issues*, *23*, 281-295.

Hu X, V Nenov, M Bergsneider, et al (2007). Estimation of hidden state variables of the intracranial system using constrained nonlinear Kalman filters. *IEEE Trans Biomed Eng* 54:597–610.

Jentoft, S and J Mikalsen (2003). A vicious circle? The dynamics of rule-making in Norwegian fisheries, Marine Policy [need vol, pgs.]

Jones A, J Homer, D Murphy, J Essein, B Milstein, D Seville (2006). Understanding diabetes population dynamics through simulation modeling and experimentation. *American Journal of Public Health*, *96*, 488-494.

Manchikanti L, K Cash, K Damron, R Manchukonda, V Pampati, C. McManus, C. D. (2006). Controlled substance abuse and illicit drug use in chronic pain patients: An evaluation of multiple variables. *Pain Physician*, *9*, 215-226.

Milstein B, J Homer, G Hirsch (2010). Analyzing national health reform strategies with a dynamic simulation model. *American Journal of Public Health*, *100*, 811-819.

Moxnes, E (1998). Overexploitation of Renewable Resources: The role of misperceptions. *Journal of Economic Behavior and Organization* 37(1):107-127.

Moxnes, E (2001). Not only the tragedy of the commons: misperceptions of feedback and policies for sustainable development, *System Dynamics Review, Vol.16*(4): 325-348.

Moxnes, E (2004). Misperceptions of basic dynamics: the case of renewable resource management, *System Dynamics Review, Vol.20*(2): 139-162.

Moxnes, E (2005). Policy sensitivity analysis: simple versus complex fishery models, *System Dynamics Review, Vol.21*(2): 123-145.

Pcouncil (2003). http://www.pcouncil.org (accessed 3/17/15)

Pilkey, O and L Pilkey-Jarvis (2007). *Useless Arithmetic. Why Environmental Scientists Can't Predict the Future,* Columbia Univ. Press.

Ruth, M and J Lindholm (1996). Dynamic modeling of multispecies fisheries for consensus building and management. *Environmental Conservation* 23:332-342.

Saysel A, and Y Barlas (2006). Model simplication and validation with indirect structure validity tests. *System Dynamics Review, Vol.22*(3): 241-262.

Schaefer, M (1954). Some Aspects of the Dynamics of Populations important to the Management of Commercial Marine Fisheries. *Bulletin of the Inter-American tropical tuna commission 1, 25-56*.

Sterman, J (2000). *Business Dynamics: Systems Thinking and Modeling for a Complex World.* Irwin McGraw-Hill.

Ursino M, C Lodi, G Russo G (2000). Cerebral hemodynamic response to CO2 tests in patients with internal carotid artery occlusion: Modeling study and in vivo validation. *J Vasc Res* 37:123–133.

Ursino M, C Lodi (1997). A simple mathematical model of the interaction between intracranial pressure and cerebral hemodynamics. *J Appl Physiol* 82:1256–1269

Ursino M, E Magosso (2001). Role of tissue hypoxia in cerebrovascular regulation: A mathematical modeling study. *Ann Biomed Eng* 29:563–574.

Ursino M, A Ter Minassian, C Lodi, et al. (2000). Cerebral hemodynamics during arterial and CO2 pressure changes: In vivo prediction by a mathematical model. *Am J Physiol Heart Circ Physiol* 279:H2439–H2455.

Van den Belt, M, L Deutsch, and A Jansson (1988). A Consensus-Based Simulation Model for Management in the Patagonia Coastal Zone. *Ecological Modeling* 110, 79-103.

Wakeland W (2007). Modeling Fishery Regulation & Compliance: A Case Study of the Yellowtail Rockfish. Proc. 25th Int'l System Dynamics Conf.

Wakeland W, R Agbeko, K Vinecore, M Peters, B Goldstein (2009). Assessing the Prediction Potential of a Computer Model of Intracranial Pressure Dynamics. *Critical Care Medicine*, Vol. 37, No. 3, pp 1079-89.

Wakeland W, O Cangur, G Rueda, and A Scholz (2003). A System Dynamics Model of the Pacific Coast Rockfish Fishery, Proc. 21st Int'l System Dynamics Conf.

Wakeland W, J Fusion J, B Goldstein (2005). Estimation of subject-specific ICP dynamics models using prospective clinical data. *In*: Modeling and Biology, IV. Ursino M, Brebbia C, Pontelli G, et al (Eds). WIT Press, S. Hampton, Boston, pp 57–66.

Wakeland W, B Goldstein B (2008). A review of physiological simulation models of intracranial pressure dynamics. *Comput Biol Med* 38:1024–1041.

Wakeland W and M Hoarfrost (2005). The Case for Thoroughly Testing Complex System Dynamic Models. Proc. 23rd Int'l System Dynamics Conf.

Wakeland W, A Nielsen, T Schmidt (2012). System Dynamics Modeling of Medical Use, Nonmedical Use and Diversion of Prescription Opioid Analgesics, Proc. 30th Int'l System Dynamics Conf.

Wakeland W, A Nielsen, T Schmidt, J Fitzgerald, J Haddox, D McCarty (2013). Modeling the Impact of Simulated Educational Interventions on the Use and Abuse of Pharmaceutical Opioids in the United States: A Report on Initial Efforts. *Health Education Behavior* 40(1), 74S-86S.

Warner, M., Chen, L. H., & Makuc, D. M. (2009). *Increase in fatal poisonings involving opioid analgesics in the United States, 1999-2006* (NCHS Data Brief, No. 22). Hyattsville, MD: U.S. Department of Health and Human Services, Centers for Disease Control and Prevention, National Center for Health Statistics. Retrieved from http://www.cdc.gov/nchs/data/databriefs/db22.pdf

Warner M, L Chen, D Makuc, R Anderson, A Miniño (2011). *Drug poisoning deaths in the United States, 1980-2008* (NCHS Data Brief, No. 81). Hyattsville, MD: U.S. Department of Health and Human Services, Centers for Disease Control and Prevention, National Center for Health Statistics. Retrieved from http://www.cdc.gov/nchs/data/databriefs/db81.pdf

White J and H Dalton (2002). Pediatric trauma: Postinjury care in the pediatric intensive care unit. *Crit Care Med* 30:S478–S488.

Yellowtail (2005). http://www.pcouncil.org/wp-content/uploads/Yellowtail_Rockfish_Final_0506.pdf (accessed 3/17/15)