Ecological Modelling 434 (2020) 109246

Contents lists available at ScienceDirect

Ecological Modelling

journal homepage: www.elsevier.com/locate/ecolmodel

Model laboratories: A quick-start guide for design of simulation experiments for dynamic systems models

Benjamin L. Turner

Department of Agriculture, Agribusiness, and Environmental Science and King Ranch® Institute for Ranch Management, Texas A&M University-Kingsville, 700 University Blvd., MSC 228, Kingsville, TX 78363, United States

ARTICLE INFO

Keywords:
Dynamic modeling
Experimental design
Sensitivity analysis
What-if analysis
Systems simulation
System dynamics

ABSTRACT

The use of dynamic systems models by scientists, managers, and policy-makers is becoming more common due to the increasingly complex nature of ecological and socio-economic problems. Unfortunately, most scientific training in the life sciences only includes dynamic modeling as elective, supplementary courses at a beginners-level, which is not conducive to generating the expertise needed to properly develop, test, and learn from dynamic modeling approaches and risks utilization of poor quality models and adoption of unreliable recommendations. The objective of this paper is to fill part of that gap, particularly regarding model experimentation, by summarizing key concepts in experimental design for simulation experiments and illustrating hands-on examples of experiments needed for developing a deeper understanding of complex, dynamic systems. The experiments include extreme conditions testing, sensitivity analyses of model behaviors given variation in both parameter values and graphical (table) functions, and "what-if?" experiments (e.g., counterfactual trajectories, boundary-adequacy tests, and intervention threshold experiments). Each experimental example describes the theoretical foundation of the test, illustrates its application using an ecological systems model, and increases in degree of difficulty from novice to advanced skill levels. By doing so, we demonstrate consistent, scientific means to glean valuable insights about the model's structure-behavior link, uncover any unforeseen model flaws or incorrect formulations, and enhance the confidence (validity) of the model for its intended use.

1. Introduction

Mathematical models, particularly dynamic systems models, are quantitative descriptions of the natural and social processes underlying the functions and patterns observed in the real world. Models have become increasingly useful for scientists, managers, and policy-makers due to their ability to capture complex natural and socio-economic processes (and the couplings between them) and present them in a way that inspires scientific creativity, improves management decision-making, informs policy-making processes, and critiques or enlightens prevailing mental models (Meadows and Robinson 1985; Sterman 1994; Sterman 2002). Despite this growing interest in and use of dynamic modeling approaches, most scientific training in the life sciences only includes modeling and simulation of such systems as elective or minor courses at a beginners-level or are applied to problems with a narrow model boundary or scope (e.g., single- to a few system processes rather than interactions between ecologic, environmental, agricultural, and economic elements), which may limit accumulation of diverse modeling expertise in such fields. The general lack of scientific training needed to generate capable expertise to properly develop, test, and learn from dynamic modeling approaches can lead to poor quality models that produce unreliable management recommendations, especially in systems that cannot be reasonably physically studied or tested because of spatial or temporal limitations. Because of this growing gap, resources are needed to aid scientists in improving their proficiency in model development and use. The aim of this paper is to fill part of that gap, particularly regarding model experimentation.

Typically, the modeling process encompasses five key steps: 1) *problem articulation*, boundary definition, identification of reference mode behaviors, establishment of relevant time horizons, and statement of modeling objectives; 2) *dynamic hypothesis formation and conceptual model development* (e.g., causal maps, subsystem diagrams, etc.); 3) *quantitative model development*, whereby equations, parameters, initial conditions, and decision rules are specified to arrive at a simulate-able model; 4) *model evaluation* (or testing), whereby developers inspect model structures and outputs to estimate its overall performance towards the model goals (which often includes comparison of model generated data to observed data, extreme condition and/or sensitivity testing, etc.); and 5) *policy/strategy design and evaluation*, whereby modelers specify scenarios of interest, often stated as "what-if?"

E-mail address: benjamin.turner@tamuk.edu.

https://doi.org/10.1016/j.ecolmodel.2020.109246
Received 8 May 2020; Received in revised form 9 July 2020; Accepted 6 August 2020

questions, and implement scientific (replicable) experiments aimed at answering the models objectives and crafting effective management interventions or policy recommendations (Sterman 2000). This process is iterative in nature where knowledge gained in one stage can be used to update and improve model components at other stages or in subsequent revision processes (Grant et al., 1997; Ford 1999; Sterman 2000).

Model testing (step 4 above) is particularly important since the results of this stage are used to evaluate model performance and behavior, validate use of the model as well as quantify the uncertainties, weaknesses, or shortcomings identified throughout the modeling process (Forrester and Senge 1980; Barlas 1989a and 1989b). Within the model testing stage, modelers typically rely on comparisons of model generated data with observed data from the real-world system. Although intuitive and illustratively simple to interpret, successful behavior mode reproduction can be a misleading indicator of model strength, since some models may reproduce similar reference modes equally well, but can create significant discrepancies in output behaviors with even miniscule changes in model structure or parameter values (Rahmandad and Sterman 2008). While it is true that no model is perfect (i.e., it cannot perfectly represent the real-world system and its behaviors), failure to fully understand the range of possible behavior modes a model may exhibit (and why it exhibits them) or to appreciate the model's weaknesses limits the learning process and bypasses opportunities for model improvement. This could become problematic, since identification of high-leverage management or policy changes (step 5 above) may not be observable if a model is lacking key information links, feedback processes, or structural elements that could have easily been included if recognized (step 4 above). Recognizing model limitations and knowledge gaps are important steps in admiring the problem at hand and appreciating its complexity, particularly in light of the complex, dynamic nature of the systems being studied, such as ecological or agricultural systems (Grant et al., 1997; Dalton 1975; Turner et al., 2016).

Unfortunately, experimental model testing is an overlooked step of the modeling process (Peterson and Eberlein 1994; Kleijnen et al., 2005), many tests that should be completed are abandoned after checking the model's ability to replicate historical data (Sterman 2002), and our intuitions about the cause-and-effect relationships, even in simple systems, is extremely poor (Sterman 1994; Cronin et al., 2009; Sterman, 2009). Therefore, it is critically important when developing and evaluating a dynamic model, one should incorporate a variety of scientific tests aimed at measuring the robustness of the model to significant parameter value or structure alterations, develop a deeper understanding of system's structure, and evaluate alternative system states arising from varying management decision rules. Learning to effectively design and implement a variety of such experiments is a useful addition to any modeling practitioner's tool box as it would aid in multiple stages of the modeling process (e.g., model boundary identification, hypothesis formation, overall model evaluation, and interpretation of key model insights) and hedge against bias or faulty inferences about model validity.

The objective of this paper is to summarize some concepts central to the design of experiments, specifically simulation experiments, and the common tests useful for developing a deeper understanding of complex, dynamic systems. This is a valuable contribution for the systems modeling fields on the whole, since much of the development of these tests reside in business management and operations research literature. To facilitate the diffusion of and improvement in using these model experiments for ecological and agricultural model applications, concepts from traditional experimental design are translated into important considerations for design of simulation experiments (Sections 2 and 3; readers interested in the theory and methodology underlying these selected simulation experiments are encouraged to start here). Then several illustrative examples are provided and discussed (Section 4; readers interested more in the application of the modeling experiments may start here and reference the theoretical sections as needed). The paper concludes with brief comments about the proper development and use of systems models.

2. Design of experiments: basic framework and key terminology

The main tenets of traditional design of experiments (DOE) include *control, replication*, and *randomization*. Experiments are investigations where the system under study is under the control of the investigator, meaning that subjects of the investigation, nature of treatments or manipulations, and measurement procedures are all set or designed by the experimenter (Cox and Reid 2000). With experimental controls, investigators should account for potential sources of error and variability through systematic DOE (Sanchez 2005). Replication (or repetition) to gain more data is required to obtain more precise results (e.g., narrower confidence intervals), while randomization (i.e., random order of experimental treatments such that one's performance does not depend on another) guards against the possible insertion of investigator biases in system response to treatments (Cavazzuti 2013; Sanchez 2005). Both replication and randomization are requirements of good DOE in order to avoid systematic errors as well as estimate the magnitude of random errors (Cox and Reid 2000).

In performing DOE, and assuming the problem at hand is well defined (modeling step 1), the experimenter chooses factors (or variables or parameters) in which to vary and the design space (or range) that each factor is allowed to vary within (Cavazzuti 2013). Factors may be qualitative, quantitative, discrete, or continuous in nature. In practice, the number and nature of parameters, treatment values or ranges, and replications are selected based on what can be afforded by the investigators. In terms of inputs (e.g., treatments), outputs (dependent variables), and goals (e.g., objective functions), the latter two are also called *response variables* while the former may be referred to as the *sample space*. A *scenario* or *design point* is a combination of values for all factors (Kleijnen et al., 2005).

Effective DOE increases the efficiency of an investigation by limiting (or eliminating) trial-and-error treatment strategies and avoiding confounding results that make implementation or adoption of findings difficult (Kleijnen et al., 2005). Some traditional DOE designs include randomized complete block, full and fractional factorial, central composite, Box-Behnken, Plackett-Burman, Taguchi, Latin Hypercube, sequential bifurcation, and frequency-based, among many others. The purpose is not to review these designs in depth but simply acknowledge the diversity of useful designs available depending on the problem at hand.[1] Below, some variations of these designs are shown in the applications specific to modeling dynamic systems. Lastly, DOE should be viewed through the context of not only the environment that an experiment takes place (e.g., field, soil, animals, etc.), but also the context of the experimenter, because no experiment exists unless someone has asked a particular question, found no suitable answer, and thought it was important enough to invest the time and resources to carry out (Pearce 1983). In the terms of dynamic systems models, experimenter context can include both the modeling or programming language one uses to carry out an experiment about the real-world, as well as the experimenter's own mental model (Peterson and Eberlein 1994).

3. Design of simulation experiments

The DOE is an essential but often overlooked step in the modeling process (specifically steps 4 and 5 defined above; Pearce 1983; Kleijnen et al., 2005; Peterson and Eberlein 1994). Many important tests are never done or investigators simply stop with replication of historical data (Sterman 2002). Still worse, unplanned, hit-or-miss

[1] Interested readers are encouraged to see references cited in this section for in-depth reviews of these procedures.

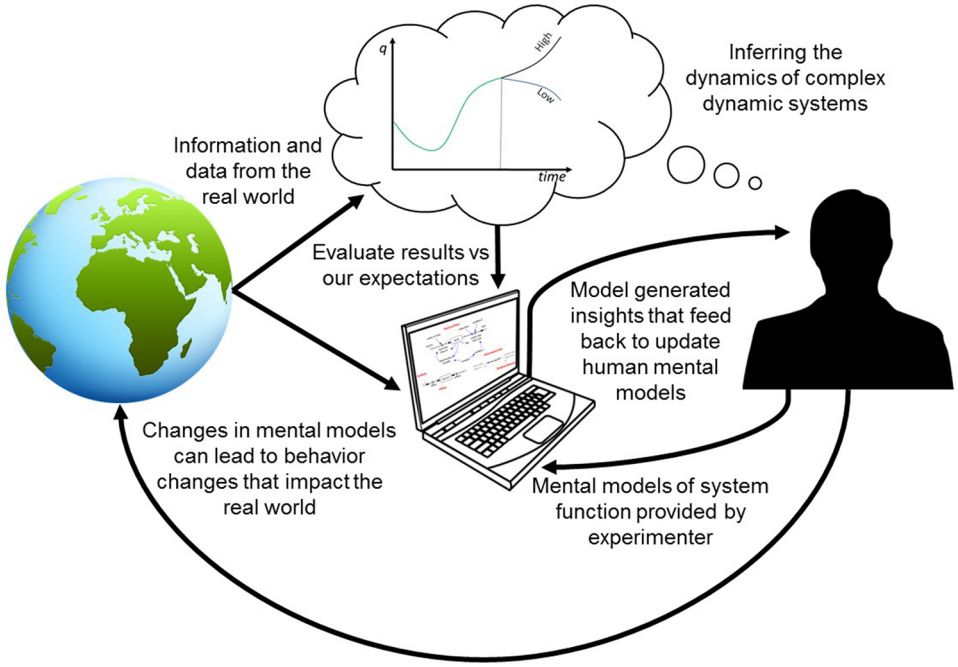

Fig. 1. Iterative management and modeling processes, whereby mental models, formal models, and the real-world outcomes feedback on each other to create the dynamics of systems we observe.

experimentation can often be frustrating, inefficient, and ultimately unhelpful (Kelton and Barton 2003). This can be extremely problematic, since the boundaries of our mental models and the inferences we make about complex dynamic systems tends to be deficient (Sterman 1994). Simulation models can be used to mimic complex systems but can be manipulated in ways that are too slow, too costly, unethical, or simply impossible to complete in the real-world (Peck 2004; Sterman 2002). For example, it may be too slow or costly to perform a grazing management study across an entire area of interest (e.g., whole ranch, county, or watershed-scales) or test the impact of redesigning an irrigation district delivery system. Likewise, when involving scarce resources (e.g., ground water or surface-water dependent systems) or threatened or endangered species, it may be viewed as unethical to proceed with the degree of manipulation needed for traditional experimentation. Models can therefore be used to represent such systems and allow for experimentation and learning about the real-world given such constraints.

Often in traditional DOE, the number of variables selected is small and their range of variability restricted due to social or economic constraints (e.g., limited time, labor, and budgets; biologic, ecologic, or ethical limitations). Modern modeling platforms overcome these experimental constraints given their ability to simulate systems across time and space rapidly and without consequences (intended or otherwise) to the real-world. Although model experiments are generally free of the constraints encountered when conducting real-world experiments, the same general principles apply, namely design of treatments and controls, randomization or estimation of uncertainty in the system, sample size considerations, and replicability so others can repeat and extend experiments elsewhere (Sterman 2000; Peck 2004; Kennedy 2019). Additionally, model experiments should be conducted in a reflective and iterative manner so that testing uncovers model flaws, challenges assumptions, and encourages critique and improvement in mental models and real-world systems (Fig. 1) (Sterman 2000). By doing so, simulation speeds up and strengthens the learning process, stimulates improvement in both mental and formal models, improves our intuition about system dynamics, and because the complex nature of dynamic systems, makes simulation the only practical way to test models (Sterman 1994; Sterman 2002).

Although there is variability in the descriptions of simulation experiments due to nuances in alternative modeling paradigms, there are at least three common shared purposes for design and use of simulation experiments: (i) evaluate the robustness of the system model, (ii) developing depth of system understanding, and (iii) comparing effectiveness of alternative assumptions, decision rules, or policies (Ford 1999; Sterman 2000; Kleijnen et al., 2005). However, it is important to recognize that there are a variety of other tests useful for model calibration and evaluation, such as behavior reproduction tests (Oliva 2003; Martinez-Moyano and Richardson 2013) that are typical of many modeling applications across skill levels. In addition, there are highly advanced model development and analysis procedures, such as bootstrapping for parameter value confidence interval estimation and hypothesis testing (purpose i and ii above; Dogan 2007), feedback loop dominance and eigenvalue elasticity analysis (purpose ii above; Oliva 2015; Oliva 2016; Naumov and Oliva 2018; Kampmann and Oliva 2020), or integration of behavioral economic theory into dynamic decision-making frameworks employed in models (purpose iii above; Langarudi and Bar-on 2018; Mohaghegh and Größler 2020) that require the most expert modeling skill and a combination of mathematical or programming software applications. Each of these are beyond the scope of this paper. Here, we focus on a non-exhaustive but comprehensive set of experimental tests to aid in the three purposes described above and that progress in difficulty to guide users from novice to advanced skill levels.[2]

3.1. Experiments for system robustness to extreme conditions

Much knowledge and information about the real-world pertains to the behaviors and consequences given extreme conditions that, if incorporated into a model, results in improved model performance both

[2] An important note here is that tests applied to dynamic models can be used to assess more than one objective. For example, an extreme condition test, which is a kind of sensitivity test, may be used to assess model equations and behavior under extreme conditions as well as to uncertain parameter values and the overall model structure (i.e., are physical laws conformed to?; are the decision rules of actors representative given extreme conditions?).

in and out of the normal operating region (Forrester and Senge 1980). Therefore, models should possess internal robustness, meaning that the behaviors produced by the model should be realistic even when extreme input, parameter, or decision-rules are imposed on the model (Sterman 2000).

Extreme condition experiments may be implemented as switches (turning variables on or off) or step or pulse functions with inordinately high or low values relative to the model's standard formulation (minus infinity, zero, plus infinity; Forrester and Senge 1980). Extreme condition tests are one of the most important experiments to consider during the model evaluation and testing stage because a) it is a powerful experiment for uncovering flaws in the model and b) it enhances model utility for analyzing how a system operates outside its normal region (Forrester and Senge 1980; Martinez-Monyano and Richardson 2013). Interpretation of extreme conditions test results depends on the application at-hand, but should follow rules of logic and reasoning regarding possible, and realistic, real-world behaviors. For example, plant production cannot occur without water for evapotranspiration, a water reservoir cannot store negative volumes of water, and small-holder livestock herd sizes cannot go below zero. Likewise, with unlimited water, plant production should reach its biological limit and then cease to grow, unlimited rainfall will fill the reservoir to its capacity at which point it begins to overflow, and livestock herd sizes may grow under favorable conditions but be subject to their available feed resources (i.e., if forage resource and stored feed levels become static herd sizes should cease to grow). The extreme condition test aids in identification of model flaws and inconsistencies, and therefore the mental models of those in the system. As such, extreme condition tests act as model "reality checks" (Peterson and Eberlien 1994).

3.2. Experiments for developing depth of system understanding

All models include some degree of uncertainty due to assumptions made about parameter values, causal relationships, the model structure itself, and errors in input data (Leinweber 1979; Hekimoğlu and Barlas 2016). Therefore, experiments are required to develop depth of understanding about the particular system of interest. There are many ways these tests can be done, including the use of step, ramp, or pulse functions as well as sensitivity analyses (Forrester and Senge, 1980; Barlas 2007). Sensitivity experiments, particularly behavior-mode (or multiple-mode sensitivity) tests, are a principal method used to shed light on the possible dynamics, distribution, and uncertainties of different behaviors that may arise from a given system (Tank-Neilson 1980; Forrester and Senge 1980; Saltelli et al., 2000). Behavior sensitivity tests identify whether or not shifts in model parameter values or relationships (including graphical or table functions) can create different behavior modes or cause the model to fail previously-passed behavior tests (Forrester and Senge 1980). The latter is an important feature of policy analysis described below.

When experimenting with uncertain parameter values, a sample design must be constructed which specifies the number of simulations to include in the sensitivity tests as well as the range assigned to each parameter value (preferably twice as large as statistical or judgmental considerations suggest; Ford and Flynn, 2005; Hekimoğlu and Barlas 2016; Sterman 2000). Although parameter values may be altered one variable at a time, it is recommended that multiple parameters be adjusted simultaneously in order to analyze their combined effect on model output (Hekimoğlu and Barlas 2016) and that any interdependencies may be identified (Ford and Flynn 2005). When conducting sensitivity experiments, it is critical to identify which type of sensitivity is being measured: numerical sensitivity (where a change in assumptions results in changes in numerical results), behavior-mode sensitivity (where a change in assumptions changes the model's patterns of behavior), or policy sensitivity (where a change in assumption reverses the impacts of a proposed decision; Sterman 2000). Each method described below outlines useful ways of measuring numerical, behavioral, and policy sensitivity.

Due to the nature of simulation in general, all models express numerical sensitivity. Therefore, one objective of sensitivity experiments is to identify which inputs or parameters have the largest effect on model output or state variables. Ford and Flynn (2005) demonstrate a statistical screening procedure for multivariate sensitivity analysis based on the correlation coefficient (CC, ranging from −1 to 1), typically denoted r, given as

$$r = \frac{\Sigma\,(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\Sigma(X_i - \bar{X})^2\Sigma(Y_i - \bar{Y})^2}} \quad (1)$$

where r represents the CC, x_i and $\bar{x}$ are the point values of the independent variable at time i and the mean x value, and y_i and $\bar{y}$ are the point values for the dependent value at time i and the mean y value. For each unit of time in the simulation, a CC is estimated for the each independent variable on the dependent variable, which can be used to identify possible interdependencies (i.e., when input variables, in the real-world, vary dependently with one another), in which model adjustments to account for relationships may be necessary. The CC also allows one to rank the most influential factors on a given state variable. Using this approach facilitates estimation of tolerance intervals with a given confidence level (Hahn and Meeker 1991; van Belle 2002).

Although sensitivity experiments on parameter values are necessary in order to understand the effect individual or groups of parameters have on a particular model, such tests must be complemented with sensitivity experiments of graphical (table) functions, which are common in dynamic systems models (Ford, 1999; Eker et al., 2014). Graphical functions represent the relationship between two variables, one the independent and one the dependent variable (Deaton and Winebrake 2000; Eker et al., 2014).[3] They are useful especially when real-world relationships between two variables are known to exist but where specific analytic equations are unknown, although if analytic equations are known from prior research these are often included as graphical functions. Because of the subjectivity associated with graphical functions there is added uncertainty in model outputs (Eker et al., 2014). Eker et al. (2014) extended a method by Hearne (2010) to generate alternative functional forms of graphical functions in an automated matter to facilitate graphical function sensitivity experiments. Based on perturbation theory, the method uses "distortion functions" multiplied by the model's original graphical function to arrive at alternative nonlinear relationships between independent and dependent variables. Distortion functions are composed of independent parameters that create alternative functional forms and are easy to test experimentally using the same procedures as those used to test parameter value sensitivity (Eker et al., 2014). The simplest distortion function, triangular, is given by

$$\mathrm{h}(x,\ p,\ m) = 1 + \frac{m\,(x - c)}{p - c}$$

$$c = \begin{cases} a, & x \le p \\ b, & x > p \end{cases} \quad (2)$$

where m represents the maximum deviation from 1 (ranging from −1 to 1) and p is the point where this deviation occurs (ranging from 0 to 1). Different combinations of distortion function parameter values result in different table function shapes. Distortion functions may be linear (triangular) or nonlinear (sine or cubic) in nature, however, triangular functions are most easily controlled and interpreted, especially using scatter plots and correlation coefficients (Eker et al., 2014).

Lastly, Hekimoğlu and Barlas (2016) demonstrate, using traditional sensitivity analysis, a means to differentiate changes in model behavior

[3] Graphical functions are also called table functions, and are often parameterized as a series of x-y coordinates representing the relationship between an input variable (x) and the response (y) that is used elsewhere in a model.

modes and quantify behavioral sensitivity outcomes. Behavior-modes are most simply defined as the pattern over time expressed in the variable of interest (see Appendix Fig. A1). The primary behavior modes are zero/constant, linear growth/decline, exponential growth/decline, goal seeking growth/decline, S-shaped growth/decline, growth and decline or decline and growth, and oscillation with/without growth/decline (Walrave 2016). When identifying shifts in behavior-mode, a behavior pattern measure must be specified and created in the model for further analysis given the lack of automated means to identify and measure behavior-mode changes. After selecting input parameters, their ranges of values, and conducting the sensitivity experiment, individual trial results are screened (by visual inspection of the output data) and grouped into those that exhibit similar behavior modes. Next, pattern measures are estimated (e.g., growth rates, periods or amplitudes of oscillations, peak-point of a boom-bust cycle, etc.) and regressed against standardized input parameter values of the sensitivity trials, i.e., by

$$\frac{x - \bar{x}}{\sigma_x} \tag{3}$$

where x represents the nonstandardized parameter value, and $\bar{x}$ and σ_x are the mean and standard deviation of parameter values.[4] Using standardized parameter values improves interpretation of regression results (Kleijnen 1995). Regression outputs can then be used to examine the nature of influence and rank of importance of each parameter on the behavior pattern, which can shed light on potential high leverage parameters in the model. For example, t-tests indicate significance while the signs of the regression coefficients indicate the direction of correlation (or polarity) between the parameter and the behavior measure (assuming parameter values used in sensitivity came from independent distributions).

3.3. Experiments for comparing appropriateness of alternative assumptions and effectiveness of new decision-rules or policies (counterfactuals or what-ifs?)

Whereas the above experiments are generally conducted by varying parameter values or functions, experiments for comparing the appropriateness of new assumptions or effectiveness of new management decisions or policies to achieve desired system behaviors are conducted through manipulating components of the model *structure* itself. Such experiments are also called intervention studies, changed-behavior-prediction tests, boundary-adequacy tests, or policy-sensitivity tests (Forrester 1961; Forrester and Senge 1980). Often these take the form of designing "what-if" experiments, which include system improvement and boundary-adequacy tests via creation of additional model structure (Forrester and Senge, 1980; Morecroft, 1988; Martinez-Moyano and Richardson 2013). What-if experiments are typically performed using ad hoc adjustments of key model parameters (e.g., Repenning, 2001; Walrave et al., 2011) as well as functional values, functional shapes, and forms of decision equations (Barlas 2007). Boundary-adequacy experiments are uniquely important because they test whether or not modification of the model boundary assumptions would change policy recommendations arrived at in the original analysis (Forrester and Senge 1980).

When future input values or forcing functions are highly uncertain, these tests may be applied retrospectively (i.e., backcasting) to analyze how "counterfactual trajectories" in key model structures would alter behaviors observed in the known past (Srinivasan 2015; Gunda et al., 2018). When inputs or forcing functions have an estimated trajectory or distribution, models can be run into the future (or past) to forecast (or backcast) potential system responses to the alternative assumptions or conditions. Once alternative scenario, decisions, or conditions are simulated, the effectiveness of proposed system changes can then be measured to identify the degree of change in model behavior as a result of the alternative assumption or scenario (Yücel and Barlas, 2015).

In order to quantify potential effects of system improvements, one may examine "what-if" questions through the analysis of intervention thresholds (i.e., the minimum intervention size and implementation time that results in the desired behavior change; Walrave 2016). Using atomic behavior patterns and their associated threshold indicators[5] (i.e., the point where a model behavior shifts from one atomic behavior pattern to another; often observed using the first and/or second derivatives), a model is iteratively tested using pre-determined intervention sizes and times, which represent the "what-if" questions of a proposed new policy, until the intervention threshold is reached or the entire search space has been simulated (i.e., where no intervention thresholds were identified). Using the resulting simulation data, an intervention threshold graph may be constructed to indicate the required intervention size at a given time to create the desired shift in behavior pattern (Walrave 2016). This approach can be implemented using sensitivity analysis (described above), but where the nature of the intervention, its size, and time applied to the model are sensitivity inputs parameters.

Lastly, given the problem-oriented nature of systems analysis, modeling generally aims to identify remedies to problems. System improvement tests seek to identify whether or not the modeling process or the policies or strategies identified by the model experiments actually led to improvements in the real-world's system structure and behavior (Sterman 2000; Martinez-Moyano and Richardson 2013). Ideally, intervention studies should follow good experimental design protocols, with control and treatment groups if possible, or with natural experiments comparing the results from those who changed behaviors as a result of the modeling process with those who did not participate (Sterman 2000; Oliva 2019).

3.4. Preparing a model for laboratory testing: notes on model calibration, evaluation, validation

Before the experimental examples are illustrated, it is important to realize that before a model is tested it must provide an adequate representation of the problem at hand. This is often captured in a *dynamic hypothesis* (DH), which is a working theory about how a system's structure of decision-rules and feedback processes generate and perpetuate the problematic behavior of interest (Richardson and Pugh, 1981; Sterman 2000). The DH should link observable patterns of behavior to micro-level structures (ecologic, environmental, socio-economic, decision-making, etc.; Forrester 1985; Morecroft 1983). A model as a laboratory should translate the DH into a quantitative working model. Therefore experimental design and testing is only as good the DH (Oliva 2003). A model calibration process should be used as a test of the DH to ensure that the model captures the observed behaviors with the right structure (Oliva 2003). There are several means in which to build confidence in a model and its DH prior to experimental testing, including hand calibration, automated calibration (see Oliva 2003 for example), and statistical evaluation to observed behaviors. Tedeschi (2006) and Bennett et al. (2013) provide reviews of statistical evaluation techniques useful for comparing model predictions with observed data from the real system. Bennet et al. (2013) also provide a general procedure for model evaluation. The critical questions one should answer during the model development and calibration phase include: 1) does the model accurately capture the DH?; 2) which data can/will be used for calibration versus evaluation?; 3) do the

[4] Standardizing is a means to rescale data (input or output) to achieve a mean of zero and standard deviation of one. Standardardization is important when variable or parameter scales differ (often the case in multivariate model sensitivity analysis) and by transforming the numerical data we add precision and stability and reduce multicollinearity issues.

[5] A summary of these are illustrated in the Appendix.

expected patterns correspond to those in the real world?; and 4) is the model precise, accurate, or both? (Tedeschi 2006). Upon successful model development, one should have adequate knowledge of the model's performance (i.e., its strengths and accuracy, weaknesses and sources of errors) to be able to properly interpret results of model hypothesis testing.

Although model validation is beyond the scope of this paper, it is a concept that all modelers must consider. Validation is the process of establishing confidence in a model such that it can be used for its particular purpose (Forrester and Senge 1980). Although some hold that model validity equates to truth of the model, many in the field of systems analysis hold that confidence is a better arbiter of validity because there is no feasible way to prove a model absolutely and completely represents reality (Forrester and Senge, 1980;Tedeschi 2006). Whether one adopts truth or confidence as their model validation criterion, it is clear that neither can be achieved without adequate model experimentation. Judging the validity of a model without having done the basic experimental tests (described below) exposes the investigator to tremendous risk, since it will be nearly impossible for them to correctly infer the full range of possible behaviors and outcomes given the dynamic nature of our ecological, agricultural, and social systems.

4. Model application: experimentation examples with discussion

To enhance the adoption, use, and documentation of simulation experiments for both novice and advanced modelers alike, we demonstrate and discuss each of the experiments described above using a moderately complex, dynamic systems model. The model was developed based on a common problem in irrigated agricultural systems that rely on surface water supplies for their water sources. In such systems, surface water flows are diverted from a river source to provide irrigation water to cropland. However, surface water diversions often has deleterious effects to native ecosystems (in terms of both habitats and individual species) through the reduction in baseline water supply that the ecosystem has relied upon for its growth and development. In addition, irrigation diversions can escalate the impact of shrinking water availability for the ecosystem depending on the season of year (which can coincide with when water use by native ecosystems is greatest), type of irrigation system used (e.g., conveyance structures for delivery and return flows as well as the type of irrigation application, such as drip, flood, sprinkler, all influence the infiltration, recharge, runoff, and return flow dynamics), and cost-effectiveness of diversion (e.g., marginal cost of water and infrastructure; level of public subsidy supporting the system, etc.). Water diversions not used for agriculture are often transferred to municipal and industrial uses that have grown in conjunction with population growth and economic expansion. Unfortunately, degradation of native ecosystems has precipitated a range of losses or reductions in ecosystem functions that support human well-being but whose economic values have not been well captured for decision-makers. The tension placed on stakeholders in these systems continue to rise due to the array economic, political, and social interests at play. The case study model profiled below captures the trade-off between agricultural water diversions from a river source and the ecologic consequences to a wildlife refuge that also relies on water from the river (Fig. 2; an expanded model illustration and supporting model files are provided the Appendix Fig. A2). The purpose of the model is to generate insight for policy-makers about the trade-offs between a surface water-supplied irrigation system and a native wildlife refuge that resides downstream from the diversion point and surrounding croplands and identify possible management strategies that balance the economic benefit of irrigation without compromising the sustainability of the wildlife refuge.

4.1. Dynamic hypothesis and model overview

The wildlife refuge is dependent on water from the river to support native ecosystem plants and the native animals that rely on those plants for survival. As water from the river source is diverted to agricultural use, river flow is reduced, which slows the growth of ecosystem plants and threatens the native animal population (Fig. 3). The model includes five state variables (crops, profits, ecosystem plants, native animals, and irrigation diversion level) and their associated rate functions for growth, decomposition or harvesting, grazing loss, reproduction and mortality, and adjustments to the irrigation diversion amounts (Fig. 2). Information links that complete the biological feedback loops include the production rate of plants, animal reproduction rates, a forage availability index (i.e., the ratio of native animals to ecosystem plants which drives the animal mortality rate), and the growth index for crops (i.e., high growth rates when total biomass is low, which slowly reduces as plants reach maturity) and associated planting and harvesting times. The major socio-economic information links include the effect of profitability on planting density and planting density on irrigation diversion rates (e.g., when profitability is enhanced, farmers raise planting density; due to the additional crop biomass, irrigation diversions are also raised to support plant growth). Additional parameters include estimates for crop water demand, irrigation efficiency, mean river flow, grazing demand of the native animals on ecosystem plants, crop prices, crop planting costs, as well as the initial investment and discount rate needed to determine the net present value (NPV) of the agricultural water use. The model, executable in the freely downloadable Vensim PLE modeling environment,[6] is provided in the Appendix material along with Microsoft Excel templates for data analysis so that readers can download and examine the model and replicate the experiments presented.

An earlier version of the model was developed and presented in Grant et al. (1997) but has been updated here to include three important feedback loops not included in the original model (two management-related feedbacks and one biophysical feedback): 1) as the net annual returns for cropping increase, the planting volume of the next year's crop also increases (positive feedback denoted 'R1' in Fig. 2), 2) as the planting volume increases, so does the irrigation diversion amount needed to support the crop (positive feedback denoted 'R2' in Fig. 2), and 3) as the number of native animals declines, pressure is applied through conservationists' effort to limit diversions from the river to agriculture (negative feedback denoted loop 'B' in Fig. 2).

To illustrate the experimental tests described above that serve to (i) evaluate system robustness, (ii) develop depth of system understanding, and (iii) compare effectiveness of alterative assumptions, decisions, or policies, we designed and subjected the model to a variety of experiments (Table 1), including extreme conditions (where the hypothesis is that the model performs in logical and reasonable ways that conform to all physical and scientific laws), sensitivity analyses of model behavior given variation in parameter values and graphical functions (where the hypotheses are that certain parameters or functions will have greater influence on the behavior of the model than others), and what-if analyses (where the hypotheses are that altering certain assumptions, decision-rules, or management strategies in the system will correct the problematic behavior). Each section that follows provides a description of how each test was implemented in the model, the results of each experiment, and how it is interpreted within the context of the purpose of the model. A summary of the rationale for each test and common obstacles encountered during implementation is provided (Table 2). These tests were chosen for several reasons: they provided a comprehensive suite of tests useful for confidence building (in both the model and the modeler), model evaluation and insight generation; they align with the "core" tests defined by Forrester and Senge (1980), they

[6] www.vensim.com/free-download/

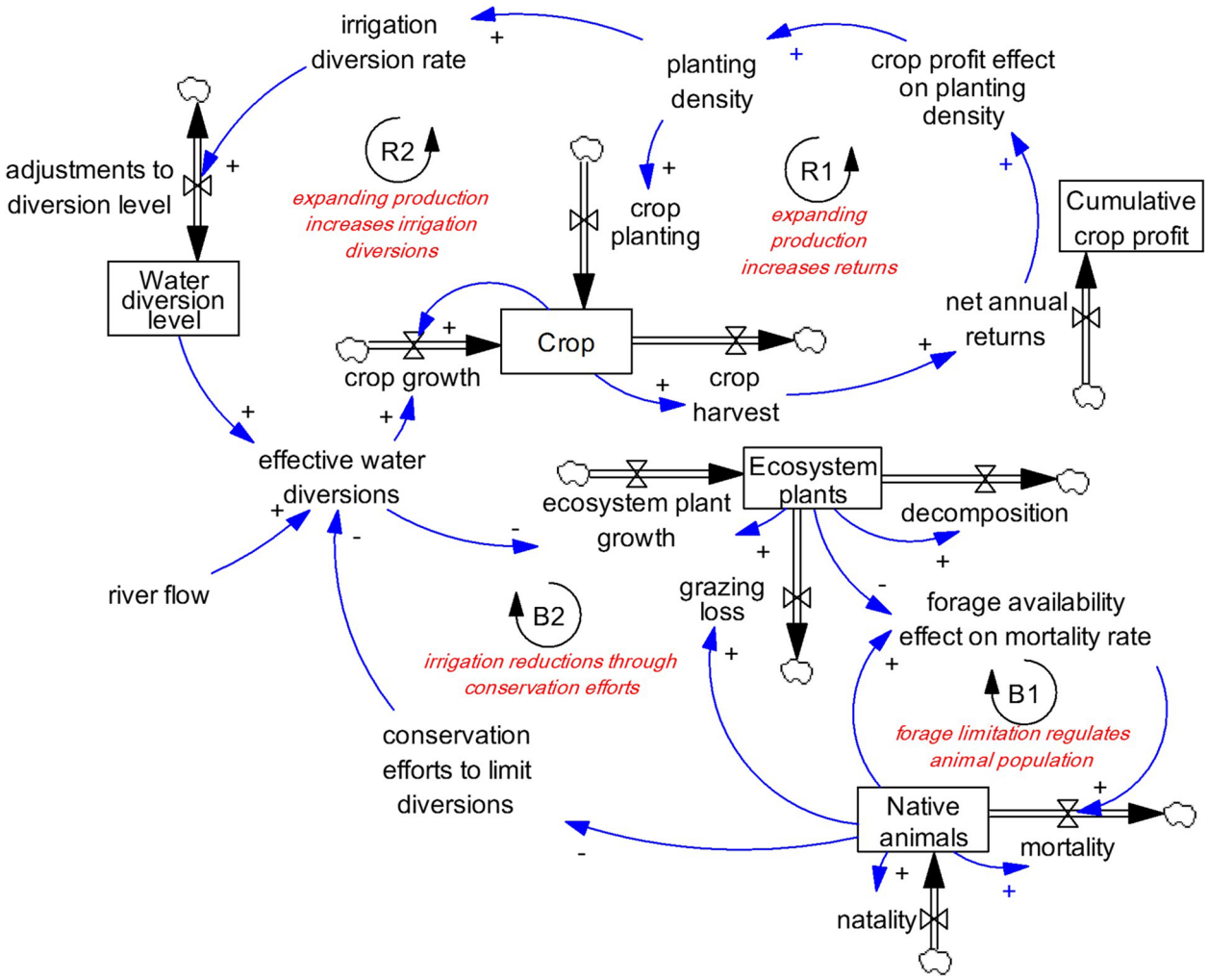

Fig. 2. Stylized stock-and-flow diagram of the irrigation-wildlife refuge case-study model. An expanded model diagram and model files can be found in the Appendix.

provided a relatively straightforward scheme moving from novice- to advanced-skill levels, and importantly, all could completed without the need for external coding or additional software outside of most dynamic modeling platforms.

4.2. Robustness to extreme conditions

System structures should permit examination of extreme combinations of states or parameter values. In order to implement extreme conditions tests, one must be familiar with the structural elements of the system (state variables, transfer functions, information converters, and variables representing time-based parameters) to be able to trace the implications of hypothetical extreme values of variables (including values known to be far outside the range of known historical or possible values) to determine the plausibility of the model's response (Forrester and Senge 1980). Here, we examine extreme conditions tests applied to several key parameters: river flow [the primary physical exogenous (i.e., arising from outside the model boundary) input to the system; calibrated to 100 c.f.s.], irrigation efficiency (the fraction of applied irrigation water that is productively consumed by the crop, taking on a dimensionless value from zero to one, calibrated to 0.5), and ecosystem plant decomposition rate (which regulates the volume of plant biomass due to metabolic costs and eventual senescence and death, taking on a dimensionless value between 0 and 1, calibrated to 0.375). Irrigation efficiency is also a significant factor influencing the surface water return flow rate back to the refuge. Here, the volume of applied irrigation water not consumed by the crop returns to the surface water source that supplies the native animal refuge. In the simplest mathematical form,

$$rfr = rf - iwd + rtf, \quad (4)$$

$$rtf = iwd - iwd*ie, \quad (5)$$

where *rfr* is the river flow to the refuge, *rf* is the upstream river flow, *iwd* is the irrigation water diverted, *rtf* is the return flow rate, and *ie* is the irrigation efficiency. Therefore, monitoring the responses to changed irrigation efficiency provides an additional check on the internal consistency of the model structure.

Abnormally high flow rates did lead to small increases in ecosystem plants earlier in the growing season, which was able to support slightly greater native animal population (Fig. 4a and b), but did not significantly alter the overall behavior patterns. On the other hand, cutting river flow to zero c.f.s. led to an eventual collapse in ecosystem plants. The stock of existing biomass became depleted by month 30, due to no new plant growth combined with monthly biomass losses from decomposition and consumption by native animals. Because of the loss in ecosystem plants, the native animal population also collapses near month 36. Importantly, neither ecosystem plants nor native animals fall below zero due to the first-order negative feedback loops (not arbitrary constraints set by the modeler) that govern the outflow of each stock (i.e., “physical inventories cannot be negative”, if the desired outflow is greater than the level of the stock, first-order negative feedback controls will regulate the outflow such that the outflow rate cannot exceed the

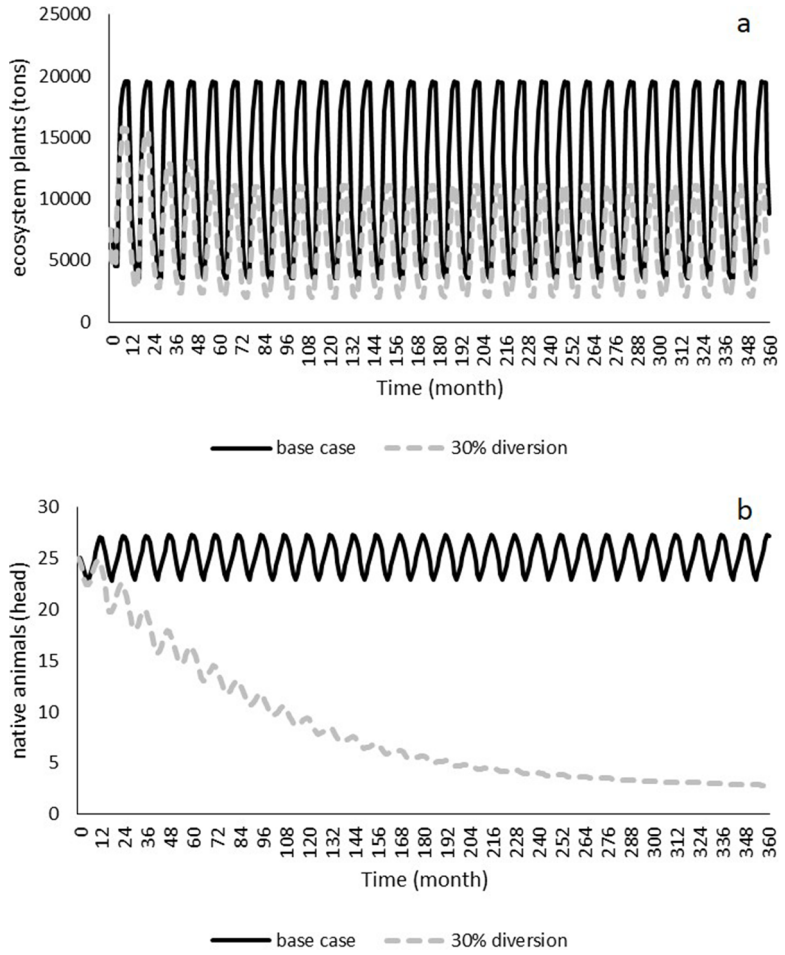

Fig. 3. Behavior-over-time graphs of ecosystem plants (a) and native animals (b) in the refuge before irrigation diversions (base-case) and after (30% of surface water diverted for agriculture) over the 30 year (360 month) time horizon of interest.

existing stock value). Consider a bathtub example. Without any inflow and the drain, or outflow, open, a full bath tub will drain water, a function of gravity and the height of the water in the tub. As the height of water declines to zero, the outflow rate also declines to zero, at which point the stock of water in the tub is empty even though the drain is open and gravitational force is still applied. Not including this kind of regulating feedback is a common pitfall among beginning modelers.

Manipulating irrigation efficiency up to 100% of applied irrigation water led to a number of significant outcomes (Fig. 4c and d). First, the 50% improvement in irrigation efficiency led to a 400% increase in crop production. Initially, this may look unreasonable. However, the modeler should examine such behaviors and be able to link them to the model structure from which they arose (Fig. 2). In this case, improving irrigation efficiency (100%) meant no water losses during irrigation (i.e., all water applied was converted into crop production). Because of the increase in production and therefore crop profitability in the first year, the strength of the two reinforcing loops in Fig. 2 (driven by profit's effect on planting density and planting density's effect on irrigation diversion levels) were greatly enhanced, leading to greater planting densities and irrigation application levels. With greater plant densities, greater water applications, and no water losses in the production system, crop production grew until reaching its maximum potential by month 36 (Fig. 4c). With irrigation efficiency at 100%, the delayed return flows to the river source were eliminated. With reduced stream flows to the refuge and the resulting loss of ecosystem plants, the native animal population declines until its complete loss by month 72 (Fig. 4d). On the other hand, when irrigation efficiency is set to 0% (i.e., no irrigation water applied is converted to crop growth, all applications become return flows), crop production was unsustainable (overlaying the x-axis in Fig. 4c) and there was no change in the native animal population (Fig. 4d), since whatever water applied was returned to the river source and supported ecosystem plant growth needed by the native animals.

The reduction in ecosystem plant decomposition rate to 0% created significant differences in ecosystem plants and native animals compared to the base case. With no decomposition, the stock of ecosystem plants was able to grow up to its biological limit subject only to losses via consumption by native animals (Fig. 4e). The native animal population, no longer subject to the seasonal variability in food supply, therefore grew in conjunction with the growth in ecosystem plants, reaching a new equilibrium population near 45 animals by month 60 (Fig. 4f). Any consumptive losses in ecosystem plants by native animals were easily compensated for by new growth.

The important insights that these experiments provide the modeler are to check that the model structure obeys basic biophysical laws. Populations and plant biomasses can't be negative, neither can they grow forever. Extreme conditions tests illuminate whether or not a model conforms to such laws. Additionally, when a model expresses a high degree of feedback (such as the feedbacks stemming from irrigation applications and their influence on return flows), such tests force the modeler to be able to explain the resulting behavior in terms of the existing model structure. When results do not conform to basic scientific laws or result in behaviors that can't be explained clearly by the existing model structure (which is a reflection of the dynamic hypothesis), the modeler should reconcile the discrepancy by revising the dynamic hypothesis, model structure, or both (Appendix Section 3.1.

Table 1
Summary and description of experimental tests along with the model variables used to facilitate model testing. Model files and data used for each test can be found in the Appendix.

Objective	Experimental test	Variables used	Units	Variable type	Test description
Evaluate robustness	Extreme conditions	River flow	c.f.s	Auxiliary	Manipulation of parameter values to extreme high and low values to ensure the model is robust enough to accommodate even the most extreme conditions.
		Irrigation efficiency	%	Auxiliary	
		Plant decomposition rate	%	Auxiliary	
Develop depth of system understanding	Step	River flow	c.f.s.	Auxiliary	Manipulation of system functions and relationships to examine behavioral changes and sensitivities in the model.
	Pulse	River flow	c.f.s.	Auxiliary	
	Ramp	River flow	c.f.s.	Auxiliary	
	Multivariate sensitivity analysis	11 simultaneous variables[#]	–	Auxiliary	
	Table function sensitivity analysis	Adjusted planting density	% of base (tons)	Graphical	
		Mortality rate	1/Mo	Graphical	
	Sensitivity of behavior pattern measures	Native animals[*]	head	Aux. of behavior pattern indicator	
Evaluate alternative assumptions, decision rules, or policies	What-if / Intervention thresholds / Boundary adequacy and changed-behavior test	Intervention times and size of expanding the wildlife refuge	tons and tons/mo. (ecosystem plants);% (expansion rate); month (expansion time)	Stocks, flows, and auxiliary	Addition, subtraction, or alteration to model structure and decision-making rules to examine the effectiveness or feasibility of new management strategies or policies.
		Construction of reservoir storage	c.f.s. (flows), acre-feet (storage)	Stocks, flows, and auxiliary	
		River flow rate	c.f.s. (flows)	Auxiliary	
	System improvement	Not well quantified	–	–	

[#] The 11 auxiliary variables in the sensitivity analysis included annual infrastructure costs, base crop planting density, base crop price per ton, base water diversions, decomposition rate, discount rate, feed resource supplement, feed resource supplement cost, irrigation efficiency, planting cost, and water consumption per ton of crop.

[*] This included native animals, mean native animals, and the first derivative of mean native animals.

provides several examples of failed extreme conditions tests often encountered in early stages by more novice modelers).

4.3. Developing depth of system understanding

To develop depth of system understanding, experiments that examine the various behavior modes that the model can create and the strength of influence that various model parameters have on creating those behavior patterns are needed. Here, we illustrate several behavioral sensitivity tests by experimenting with alternative parameter values or functions (including graphical functions) to analyze their influence on the resulting model behavior. These include varying parameter behaviors rather than their values, multivariate sensitivity analysis, graphical function sensitivity analysis, and analysis of behavior-modes (i.e., behavior pattern measures).

4.3.1. Behavioral sensitivity to model parameters

The first experimental examples demonstrate changing single parameters in unique ways via giving them dynamic rather than static values over the course of a simulation. Because of the importance of river flow as the primary water input for the ecosystem plants and native animals in the refuge as well as the crop production system (Fig. 2), we illustrate three experimental tests of the model using step, pulse, and ramp functions to vary the river flow input values into the system. Each of these tests are also applicable to other parts of the system (e.g., crop production, economics, and ecosystem components), but due to space considerations only tests of river flow are presented here.

First, using a step function, we implement a *step volume* change to river flow, at a given *step time*, such that

$$\text{river flow} = \begin{cases} 100, & t < \textit{step time} \\ 100 + \textit{step volume}, & t \geq \textit{step time} \end{cases} \quad (6)$$

where t represents time, *step time* = 120 months, and *step volume* is equal to 50 c.f.s. (but is adjustable based on user input for smaller or larger step changes, including potential reductions in river flow for negative values of *step volume*).[7]

The resulting behaviors of the river flow step experiment illustrate several immediate and delayed responses to changes in flow behavior (Fig. 5a). Due to the immediate increase in available water (150% of base river flow) and no delay in irrigation diversion and application during the growing season, crop production was enhanced nearly 50% (Fig. 5b). In a similar fashion, ecosystem plants immediately recovered from the reduction created by diversions in the base case river flow and actually reach a new peak in primary productivity (approximately 20,000 tons compared to 15,000 tons, or an increase of 33%, prior to irrigation diversion; Fig. 5c).

[7] In the Vensim modeling environment, this may be achieved a number of ways but is best implemented with use of the step function [STEP(step volume, step time)].

Table 2
Summary of the purpose and rationale for each experimental test and description of common obstacles, barriers, or limitations encountered during experimentation that potentially limit use of test (indicated from novice to advanced modelers).

Experimental test	Purpose and rationale	Common obstacles, barriers, or limitations encountered
Extreme conditions	•Does the model structure withstand extreme conditions such that the resulting behaviors are physically realistic (e.g., non-negative physical stocks; growth processes that can't grow forever)? •If not, what internal structural elements need to be added (e.g., first-order negative feedback on physical outflows) or revised (e.g., internal decision rules)?	•Improper model formulation on flow or rate variables (for novice modelers) •Identification of useful parameters to stress to extreme conditions (for novice modelers) •Requires extensive familiarity with the real-world system (for novice modelers) •Low level of insight (relative to other experiments) given the time investment made to generate test and analyze results (for advanced modelers)
Step, pulse, and ramp functions	•Does the model respond to changes in input parameter behaviors in realistic and explainable ways? •In what ways do the behavior mode of input parameters influence the behavior mode of model's endogenous structure?	•Improper model formulation on flow/rate and/or auxiliary variables (novice) •Identification of useful parameters for testing as well as the nature and degree of change (novice) •Interpretation of model behavior changes can become more difficult the farther one moves away from the test variable (novice to advanced)
Graphical sensitivity analysis	•Is the model over-sensitive to the form of graphical (table) function form? Do numerical assumptions underlying graphical functions create significantly different behavior patterns?	•Graphical functions incorrectly parameterized (novice modelers) •Significant number of auxiliary variables are needed for experimentation (novice to advanced)
Statistical screening and behavior pattern measures	•Which of the hypothesized management levers have the greatest numerical impact on the system variables of interest? Have we distinguished the "critical few" variables from the "insensitive many"? •Which of the hypothesized management levers have the greatest impact on creating an alternative behavior pattern in the system variables of interest?	•Identification of variables to be included in sensitivity simulations (novice) •Ability to identify and differentiate between different behavior modes (novice) •Determination of range of values that variables can take on during simulations (novice to advanced) •Significant number of auxiliary variables may be needed to capture behavior pattern measures needed for experimentation (novice to advanced) •Statistical screening becomes increasingly labor intensive as one moves beyond one output variable of interest (novice to advanced)
Counterfactual trajectory analysis	•What if some of the key underlying assumptions used to build the model wrong? How might the model behave if these assumptions were altered to reflect counterfactual scenarios?	•Identification of core assumptions underpinning model structure (novice) •Generating plausible alternative assumptions for identified variables (novice to advanced)
Model boundary adequacy test	•How can we inform stakeholders about the implications of policies/strategies not yet considered because they are likely to arise outside the existing model boundary? Would pursuing such a strategy alter the recommendations made from the modeling project?	•Ability to distinguish model structure necessary for the purpose versus structure that is not needed or confounding (novice) •Ability to identify and differentiate between different behavior modes (novice) •High degree of modeling skill required to conceptualize new model structures (novice to advanced)
Intervention thresholds analysis	•How much investment should be made and when in order to get the desired behavior pattern in the system variable of interest?	•Ability to identify and differentiate between different behavior modes (novice) •High degree of modeling skill (novice to advanced) •Computationally demanding (novice to advanced)

The reason for the disproportionate response in crop production and ecosystem plants are due to the disproportionate change in total inflows received by each area. Recall that the base river inflow was 100 c.f.s., irrigation diversions were 30% of the flow, irrigation efficiency was 50%, and water not consumed by the crops became return flow back to the river source. In response to the step change in river flow, the total inflow available for crop production increased 67% (i.e., $\frac{\text{new waer available} - \text{base water available}}{\text{base water available}}$; or $\frac{(50\text{ cfs} * 50\%) - (30\text{ cfs} * 50\%)}{(30\text{ cfs} * 50\%)} = \frac{25\text{ cfs} - 15\text{ cfs}}{15\text{ cfs}} = 66.7\%$). On the other hand, total inflows to the refuge only increased 47% (i.e., $\frac{125\text{ cfs} - 85\text{ cfs}}{85\text{ cfs}} = \frac{40\text{ cfs}}{85\text{ cfs}} = 47\%$). Finally, the native animal population does recover to its equilibrium level of 25 head, albeit with a delay of 160 months (Fig. 5d). Unlike the immediate plant response to water observed in crop production and ecosystem plants (which are subject to delays of less than one year in their growth capacities), native animals are subject to biological delays for gestation and birth which lengthen the recovery time for the general population. Because of this delay, the recovery in native animals is not a sharp, one-point-in-time increase, but an s-shaped growth pattern from the reduced state back to the natural population level.

In a similar fashion, we can utilize a pulse function to periodically increase or decrease a particular rate or parameter values at a desired interval (*pulse time*). In this case, a pulse is used to reduce river flow by 50% during the ecosystem plant's growing season in three year intervals to mimic the natural occurrence of drought conditions (the base case without irrigation is used here in order to more effectively observe what changes, if any, occur in native animals, which may not be perceptible given the influence of irrigation diversions shown in previous sections). The droughts begin at the start of the growing season (month of year = 4) midway through the simulation (t = 180 months), which

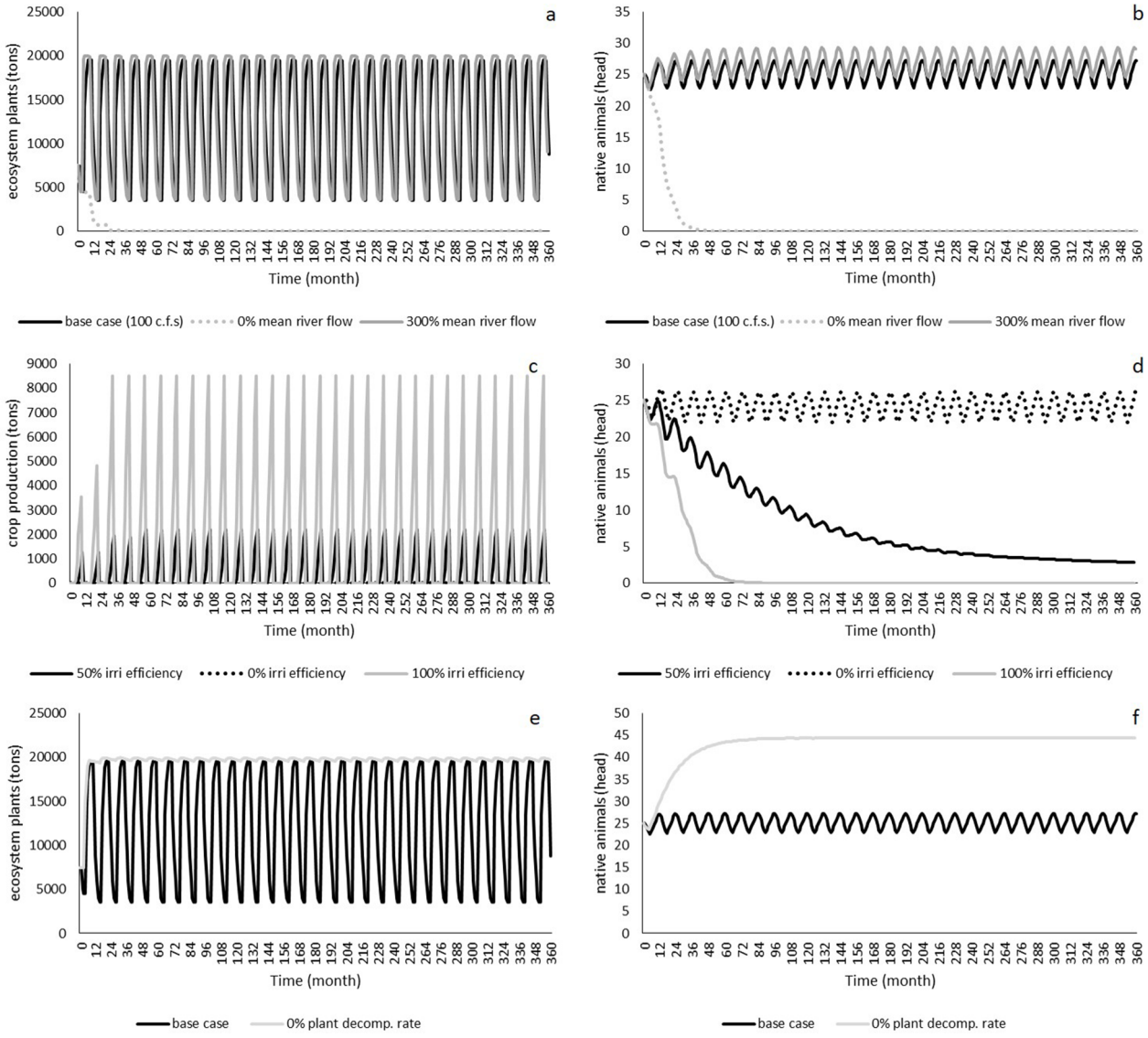

Fig. 4. Behavior-over-time graphs illustrating the resulting behaviors of extreme conditions tests of river flow (panels a and b; base case 100 c.f.s, 0 c.f.s, and 300 c.f.s), irrigation efficiency (panels c and d; base case 50%, 0%, and 100%), and ecosystem plant decomposition rate (panels e and f; base case 0.375% month^{-1}, 0% month^{-1}).

mathematically can be expressed as

$$\text{river flow} = \begin{cases} 100, & t < pst \\ 100 + pv, & t \geq pst \end{cases} \tag{7}$$

and

$$\text{pv} = \left[\frac{\tau}{T} + \sum_{n=1}^{\infty} \frac{2}{n\pi} \sin\left(\frac{n\pi\tau}{T}\right) \cos\left(\frac{2\pi n}{T} * \left(pst - \frac{\tau}{2}\right)\right) \right] * dc \tag{8}$$

where *pv* is the pulse volume, *t* is time, *pst* is the pulse start time equal to 184 months (i.e., month 180 being the start of the 16th year, plus four months to arrive at the start of the natural growing season), pulse time τ is 6 months, and pulse *period* T is 36 months. The pulse in the above function produces a sequence of re-occurring events (representing drought) via the use of a square wave function, where the pulse width is 6 months, the space width is 30 months, and the cycle time is 36 (i.e., duty cycle = 16.67%). The square wave functions yields a value of 1 when the wave is positive or 0 when it is not, thus, in order to simulate the pulse of a particular volume, the square wave is multiplied by an auxiliary variable representing the desired change, *dc* (i.e., in this case −50 c.f.s.; Fig. 6a).[8]

Similar to the step experiment above, the pulse experiment reducing river flow by 50% yielded both immediate and delayed responses. Clearly in the years with drought ecosystem plants were negatively affected, but do recover in subsequent years following each drought (Fig. 6b). Native animals are also negatively affected, but due to the biological delays inherent in the population, drought effects in one year compound into the future. For example, the reduction in ecosystem plants during the first drought pulse results in diminishing the native animals nearly 50%, but the plants are able to fully recover in the following years before the next drought occurs, whereas native animals only improve to a mean of 17.5 animals (a 30% reduction for their starting population). Because of the lengthened recovery time in native animals, the effect of each subsequent drought is amplified and reinforces the native animal population to a new, albeit lower, equilibrium level (Fig. 6c). The result is a change in behavior pattern from a

[8] In the Vensim modeling environment, this can be simulated via the pulse train function [PULSE TRAIN (start time, pulse width, cycle time, end time)*-desired pulse volume].

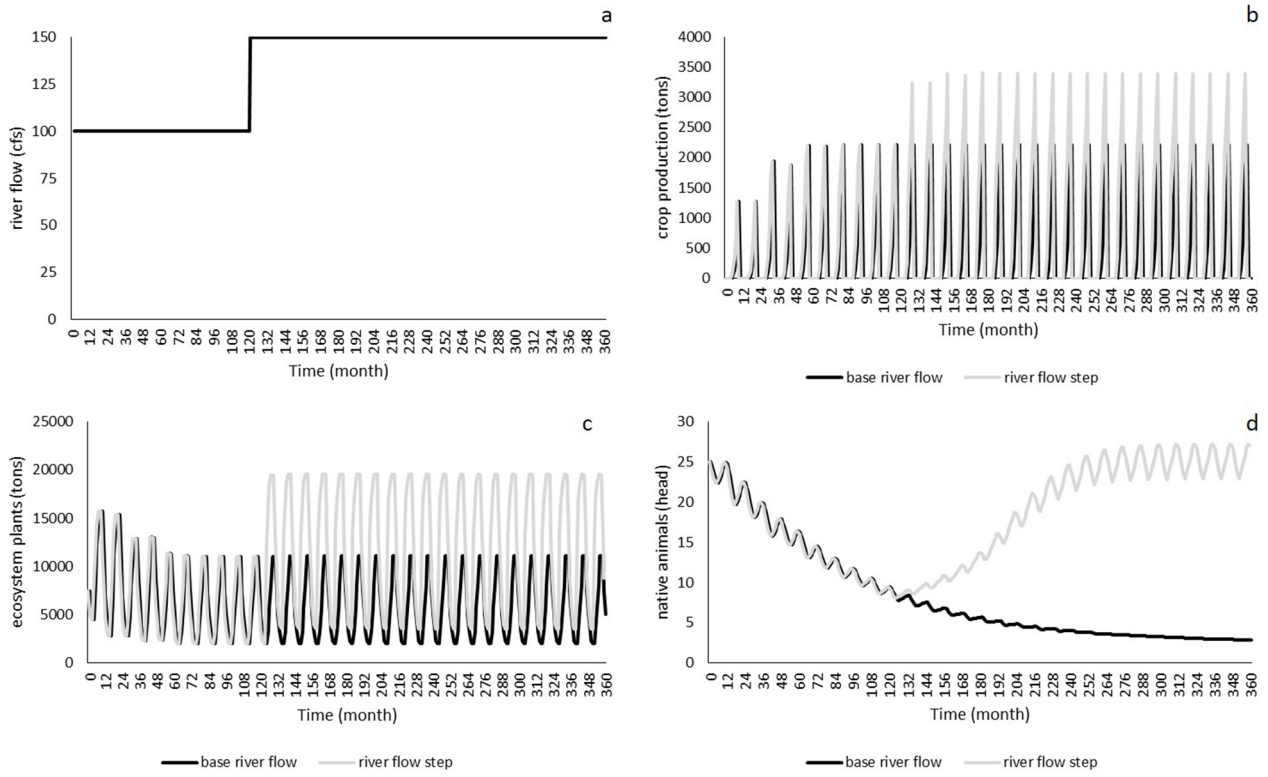

Fig. 5. Behavior-over-time graphs illustrating the resultant behavior changes caused by a step in river flow at 120 months (panel a) to crop production (panel b), ecosystem plants (panel c), and native animals (panel d).

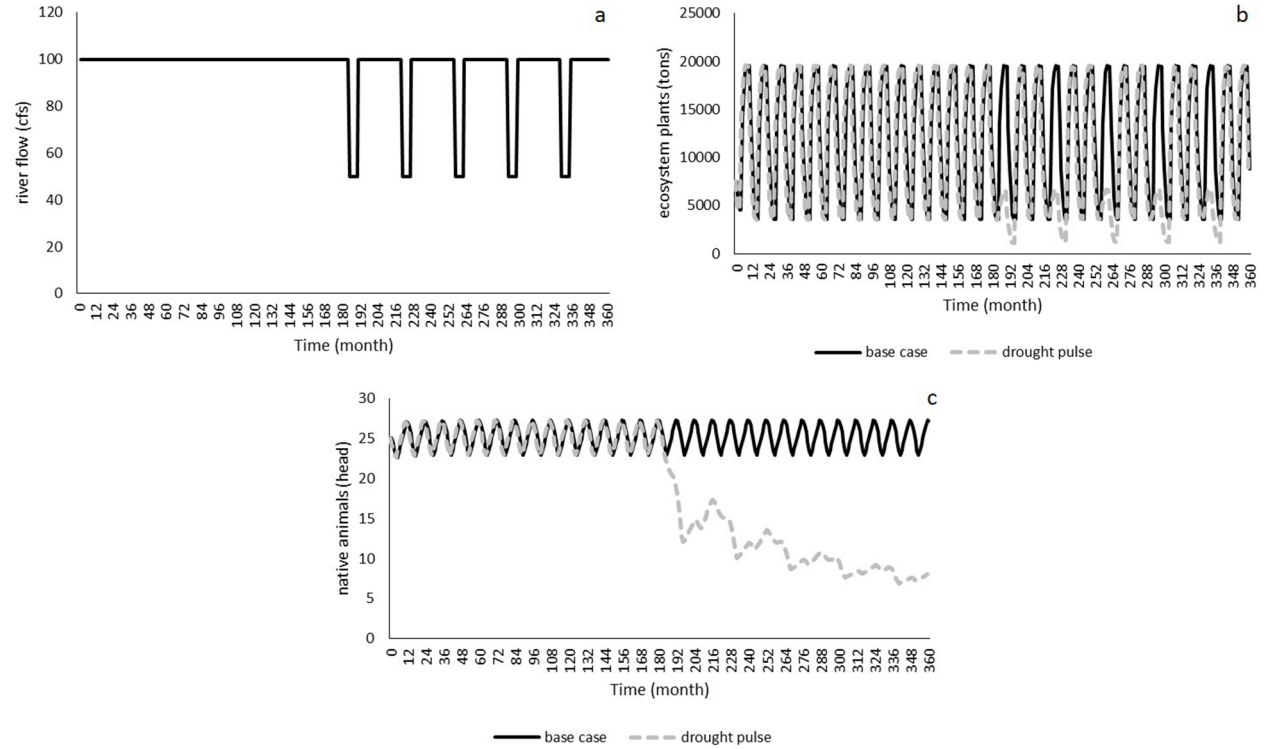

Fig. 6. Behavior-over-time graphs illustrating the resultant behavior changes caused by the −50 c.f.s. pulse series in river flows beginning at 160 months (panel a) to ecosystem plants (panel b), and native animals (panel c).

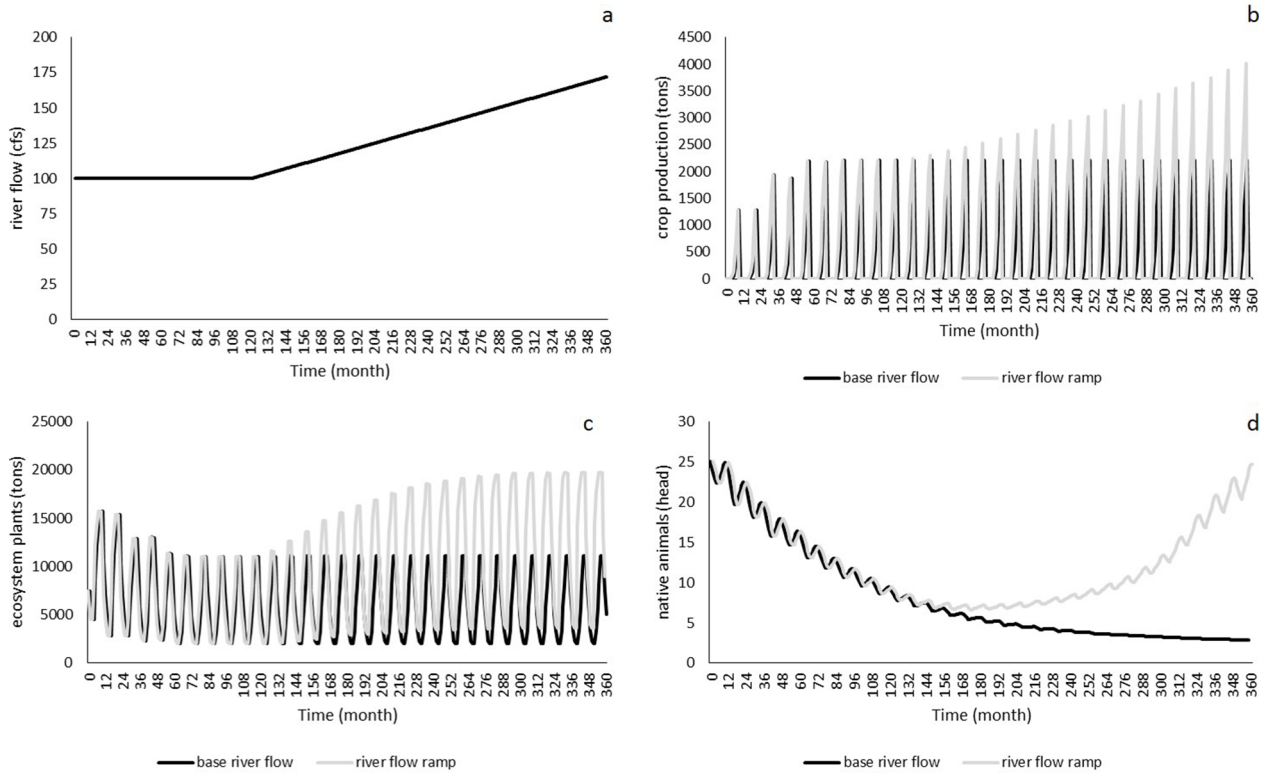

Fig. 7. Behavior-over-time graphs illustrating the resultant behavior changes caused by a ramp in river flow at 120 months (panel a) to crop production (panel b), ecosystem plants (panel c), and native animals (panel d).

sustained oscillatory behavior to one that is a declining, goal-seeking, damped oscillation.

The third test demonstrates a ramp function, which alters a parameter value from a static value to one that takes on a particular slope (positive or negative) for a particular time-span. In this case, river flow increases 0.3 c.f.s. per unit of time beginning at 120 months. Mathematically, this is expressed as[9]

$$\text{river flow} = \begin{cases} 100, & t < \textit{ramp time} \\ 100 + (\textit{ramp slope}^*(t - \textit{ramp time})), & t \geq \textit{ramp time} \end{cases} \quad (9)$$

where t is time, *ramp time* is 120 months, and *ramp slope* is 0.3.

Given the above *ramp slope* and time inputs, river flow increased from 100 c.f.s. at month 120 to 175 c.f.s. by month 360 (Fig. 7a). Due to the gradual increase in available water, crop production grew linearly in step with river flow (Fig. 7b), while ecosystem plants exhibited more of goal-seeking growth behavior up to its maximum productive potential near 20,000 tons (Fig. 7c). Because the recovery delay of ecosystem plants takes around 180 months (from 120 to 300 in Fig. 7) and the reproductive delays of native animals (described above), the native animal population exhibits an exponential growth pattern, where each subsequent gain in population, although initially small, compounds into the future (Fig. 7d). However, since ecosystem plants do reach a point where growth no longer occurs, a reasonable follow-up test would be to extend the simulation out beyond 360 months to identify at what point (if any) the native animal and crop production behavior patterns shift from exponential and linear growth to some other behavior patterns, most likely goal-seeking.

Tests of these types aid the modeler in understanding how model structure produces unique behavior modes given a dynamic input. These are especially valuable for management or policy-related questions due to the fact that most dynamic models have the ability to exhibit multiple behavior-modes, which makes it possible to examine potential interactions between modes or how different decision-rules lead to particular behavior patterns (Forrester and Senge 1980; Appendix Section 3.2 provides examples of failed step, pulse, and ramp experiments that would necessitate mental and quantitative model revision).

4.3.2. Sensitivity analysis with graphical functions

The second set of experiments illustrate graphical function sensitivity analyses. In this case, two key graphical functions warranted investigation: *native animal mortality rate* (a graphical function based on per capita forage availability effect on mortality; B1 in Fig. 2) and *crop profit effect on planting density* (R1 in Fig. 2). Because mortality rate directly influences the native animal population and does not feed back to the cropping system, we only examine the effect of altered mortality rates on the animal population. On the other hand, the influence of expected planting density reaches beyond the cropping system to second- and third-order effects on the ecosystem plants and native animals; therefore, we examine crop system financial performance (via the net present value, or NPV, of irrigation, which integrates crop production and economic prices and costs) as well as the native animal population. Each of these graphical functions were manipulated using the distortion function procedure outlined by Eker et al. (2014) after Hearne (2010), illustrated with two simple examples in Fig. 8. Although there are an array of alternative distortion functions one may use to manipulate the graphical function, here we employ the one of the simplest, a single point triangular distortion (Eq. (2)), in order to minimize additional model variables required for the experiment while maintaining a distortion that is easily interpretable (for a full discussion on the strengths and weaknesses of variable distortion function possibilities, see Eker et al., 2014).

To complete the sensitivity analysis, 100 simulations were

[9] In the Vensim modeling environment, this function is most easily implemented use the ramp function [RAMP(ramp slope, ramp time, end time)])

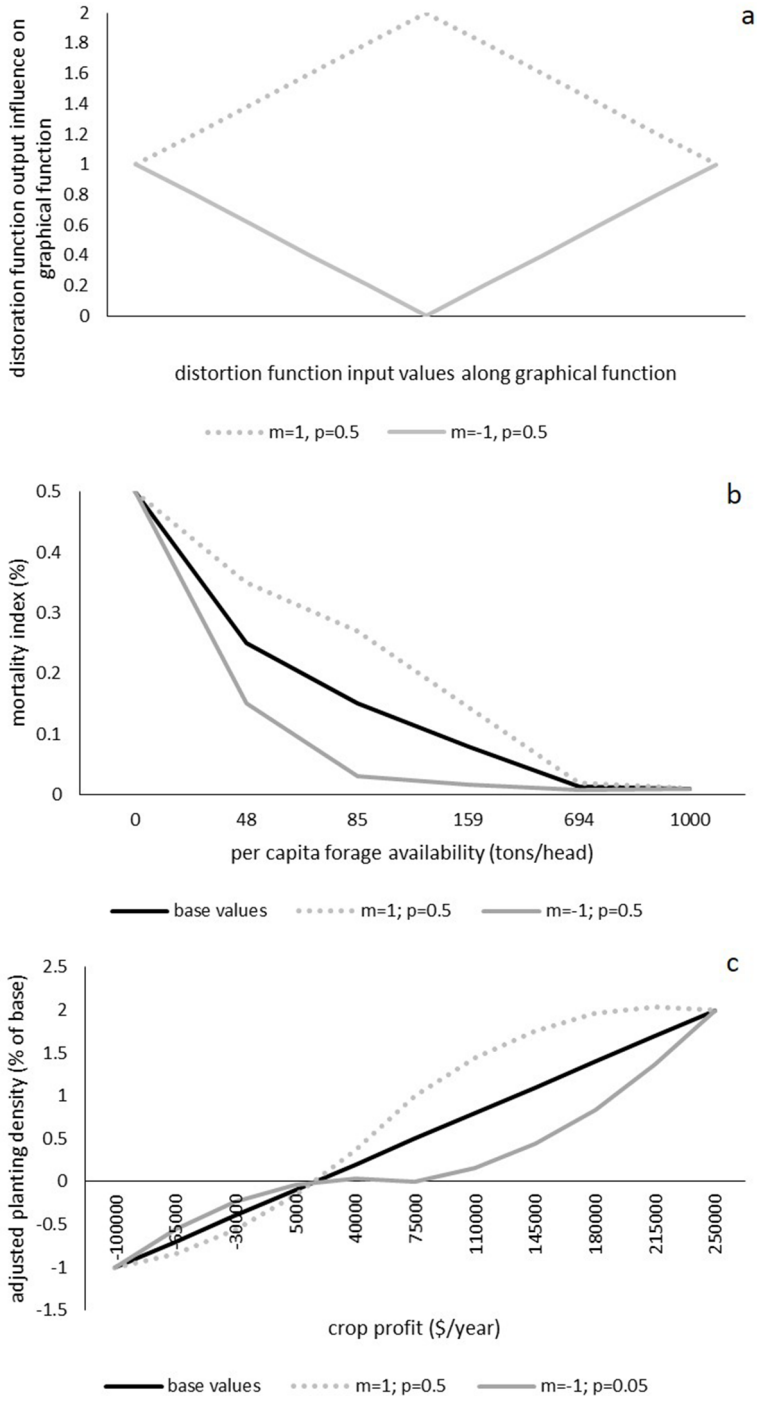

Fig. 8. Visual illustration of the effect that various possible distortion functions (panel a) contort existing graphical functions (panels b and c). Factor *m* represents the maximum departure or deviation from the initial graphical function (ranging from −1 to 1) while *p* is the point where this deviation occurs (ranging from 0 to 1, zero and one are the end points of the graphical function).

completed for each graphical function by varying the maximum deviation, *m*, from −1 to 1, and point of maximum departure, *p*, from 0 to 1 (see Eq. (2)).[10] For the most part, the model was insensitive to changes in graphical functions (Fig. 9). Altering the mortality rate function revealed primarily only one behavior pattern, decline-and-goal seeking (Fig. 9a), with only three out of the 100 simulations producing a markedly different ending population size. Examining the native animal mean and standard deviations showed that the monthly variation in population was demonstrably low relative to changes in the mortality rate graphical function (Fig. 9b). Similar results were observed for changes in expected planting density (Fig. 9c and d). Only two out of 100 simulations produced a significantly different outcome for both native animals and NPV of irrigation (Fig. 9e and f). Following Ford and Flynn (2005), correlation coefficients (CC) were calculated between the inputs, *m* and *p*, with the observed output data in native animals and NPV of irrigation (ending CC values shown in Table 3). Increasing the magnitude of the departure, *m*, from the original graphical functions for

[10] Tests were implemented in Vensim modeling environment using the built-in Monte Carlo simulation feature, with the following specifications: number of simulations 100, noise seed 1234, Latin Hypercube sampling, with random uniform distribution of input parameter values.

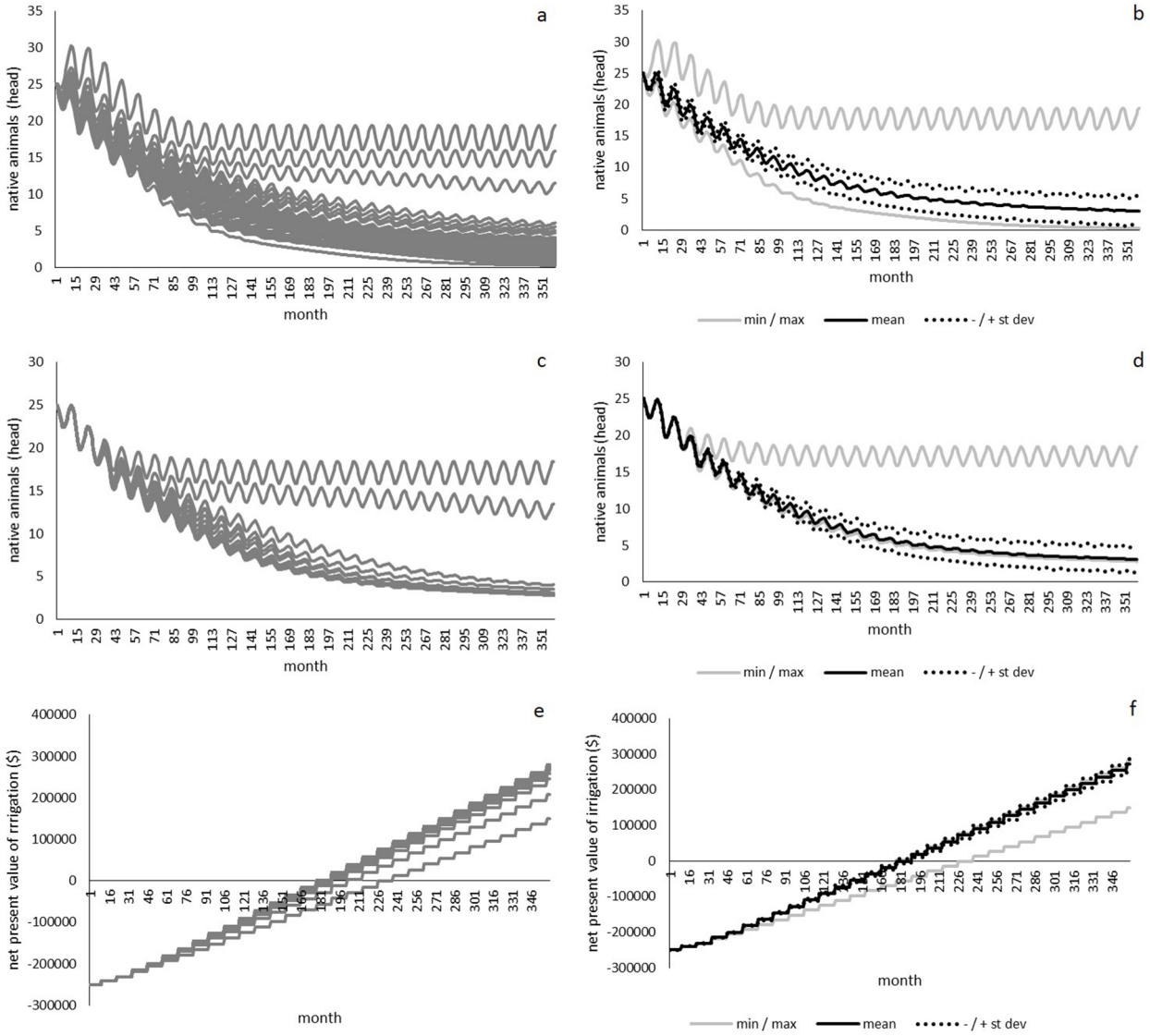

Fig. 9. Results of graphical function sensitivity analysis given 100 simulations of alternative graphical functions: all simulation runs of native animals (panel a) and the mean, minimum, maximum, and standard deviations (panel b) from altering animal mortality rate; all simulation runs of native animals (panel c), native animal mean, minimum, maximum, and standard deviation (panel d), all simulation runs of net present value of irrigation (panel e), and the net present value mean, minimum, maximum, and standard deviation (panel f) from altering expected planting density.

Table 3
Correlation coefficients between maximum departure magnitude, m, and point of maximum departure, p, used in graphical function sensitivity analysis with native animals and the net present value (NPV) of irrigation.

		Output measures	
Graphical function	input	Native animals	NPV of irrigation
mortality rate	m-factor	−0.47	–
	p-factor	−0.18	–
expected planting density	m-factor	−0.18	0.06
	p-factor	−0.25	0.33

mortality rate resulted in smaller native animals populations (indicated by negative polarity in the sign of the CC). Increasing the point of departure, *p*, also had a negative but weaker influence on native animals. This may also be intuited given the slope of the graphical function (Fig. 8b) given that larger departure values on per capita forage availability will raise mortality rate (i.e., smaller per capita forage availability → larger mortality rate → smaller native animal population). Similar effects on native animals were observed given *m* and *p* distortions on expected planting density. Conversely, larger departures had a positive influence on the NPV of irrigation (i.e., greater planting density → greater crop production and harvest → greater crop profit → greater NPV).

Graphical functions and their particular shapes are often the most debated functions during model development and overlooked parameter values during model testing. Given the extraordinary range of possibilities that graphical functions can take on, it is important that the model withstand alternative graphical forms. All models express numerical sensitivity to changes in parameter values, but models should be able to produce the behavior mode described in the dynamic hypothesis for a wide range of parameter values (including graphical functions).

4.3.3. Sensitivity of behavior patterns

The final sensitivity analysis demonstrated here pertains to multivariate tests using statistical screening (Ford and Flynn 2005) and behavior pattern measures (Hekimoğlu and Barlas 2016). To facilitate both tests, one set of 11 parameters and their ranges were specified (Table 4) and simulated for 100 model runs (a form of Monte Carlo

Table 4
Model parameters, their base values, and bounds on the range of uncertainty applied to the 100 model simulations during statistical screening and behavior pattern sensitivity analyses.

Model parameters	Base value	Lower bound	Upper bound
Agriculture			
Base crop planting density (ton/unit area)	100	75	125
Base water diversions (% of river flow)	30%	20%	40%
Irrigation efficiency (% water applied converted into crop production)	50%	40%	60%
Water consumption per ton of crop (c.f.s./ton)	0.3	0.2	0.4
Ecologic			
Decomposition rate of ecosystem plants (%)	37.5%	35%	40%
Feed resource supplement (tons/head/month)	0	0	0.25
Economic			
Crop price ($/ton)	$30	$20	$40
Planting cost ($/ton)	$50	$25	$75
Annual infrastructure cost ($/year)	$10,000	$5,000	$15,000
Feed resource supplement cost ($/ton)	$295	$215	$375
Discount rate (%)	3%	2%	4%

simulation) to observe the resulting behaviors in key system variables: native animals, ecosystem plants, total crop harvest, the NPV of irrigation, and several key behavior pattern measures for native animals (described below). Native animals, ecosystem plants, and total crop harvest were chosen because these were the key stocks of the model, that, in the real-world, are most likely to be measured and monitored given the problem at hand, while NPV of irrigation (an auxiliary variable) was chosen given that it integrates the outcomes that arise from the interrelated nature of the model's agriculture, ecologic, and economic components (Table 4). Unfortunately, comprehensive sensitivity analysis of all uncertain parameters over their entire range of possible values is for most practical purposes impossible (Sterman 2000). Given this constraint, parameters used for sensitivity testing must be selected for, typically by identifying: those you suspect (hypothesize) are both highly uncertain and likely to be influential, those that are not entirely under control of decision-makers but must be managed if a desired outcome is to be achieved, which parameters are deemed most important to decision-making by working with stakeholders or problem-owners, defining “worst” and “best” case scenarios about the problem at hand and then backing into variables needed to create such scenarios, or some combination of the above (Sterman 2000). In this particular case, sensitivity input variables were selected based on their hypothesized uncertainty and influence on the system and those that, although not entirely within control of stakeholders, are key factors that must be managed if a solution is to be reached (Table 4).

The statistical screening approach was used to evaluate the relative strength and polarity of each parameter on the key variable output behaviors, where the input and output values are regressed for each time-unit of the simulation. Varying these input parameter values simultaneously created much more variability in system behavior-modes (Fig. 10) relative to the previous sensitivity tests described above. First, both total crop harvest and NPV of irrigation (which express linear growth behavior-modes) indicated a clear shift or break between profitable and unprofitable combinations of inputs (Fig. 10a and c). However, in general, both total harvest and NPV of irrigation grew linearly (Fig. 10b and d). On the other hand, ecosystem plants and native animals expressed several behavior modes (Fig. 10e and g), in general exhibiting a goal-seeking decline behavior-mode (Fig. 10f and h).

Examination of the ending CC values between input parameters and system stocks indicate the relative strength and polarity of the link from inputs to model behavior. Positive CC values indicate positive polarity, negative CC values indicate negative polarity, while values closer to 0 indicate relative weaker degree of influence and values closer to -1 or 1 indicating stronger degree of influence on the system behavior. The two variables with the strongest influence across the entire system were base water diversions and irrigation efficiency (CCs ranging from -0.76 to $+0.58$; Table 5). The next most influential variables included base crop planting density and price per ton of crop (CCs ranging from -0.29 to $+0.27$). The remaining input parameters held relatively weak influence on system behaviors (CC values between -0.15 and 0.15). A key insight of the CC analysis that corroborates the dynamic hypothesis as well as previous model testing is the polarity between agricultural components (total crop harvest) and the ecosystem (ecosystem plants and native animals). For the most part, parameters that have a positive-polarity influence on agricultural components expressed a negative-polarity influence on ecosystem components, and vice versa. The value of the CC analysis is that it illustrates and quantifies the trade-offs between the agricultural and ecological components in this system. Lastly, one may notice the annual cycle in ecosystem plants and native animals, which is driven by the annual growth season in plants and the resulting balance with native animals. However, the trajectory of their behaviors does not express this annual oscillation, since the overall behavior pattern changes depending on whether or not the growth was greater than or equal to decomposition and grazing (in which case plants are stable, if not, they decline), or if births were greater than or equal to deaths (in which case animals are stable, if not, animals decline). Therefore in this case, the CC values would not be significantly different if the cycle point varies at the end of the simulation (i.e., the trajectory or spread in CC values would not be different, only the ending point values). However, in more cyclical or oscillatory systems, one may examine the behavior pattern of the CC itself, to understand how the CC between inputs and behavior patterns evolve over the course of the simulation (Ford and Flynn 2005).

To take statistical screening a step further and complete the behavior pattern measures analysis, the same parameters and ranges of values were used but rather than observing changes in specific system variables, we observe the resulting changes in the output of behavior pattern measures, in this case for native animals. After visually screening the sensitivity results, we observe that native animals expressed several behavior modes: 1) goal-seeking decline-and growth to a new equilibrium; 2) goal-seeking decline to a reduced population; 3) goal-seeking collapse (defined as population-levels below three animals, which effectively eliminates any possibility of successful species reproduction); and 4) linear decline (Fig. 10g and h; generic behavior modes shown in the Appendix Fig. A1). After grouping the data based on visually classifying the behavior patterns, we specify pattern measures that differentiate each unique pattern: equilibrium level (pattern 1, $n = 39$) defined as the mean native animals at the final month; inflection point and inflection level (pattern 2, $n = 12$) defined by the largest value of the first derivative of mean animals and the level of animals at that point; time to collapse (pattern 3, $n = 43$) defined at the time at which native animals falls below three, and slope (pattern 4, $n = 6$) defined as the difference between initial and final native animals

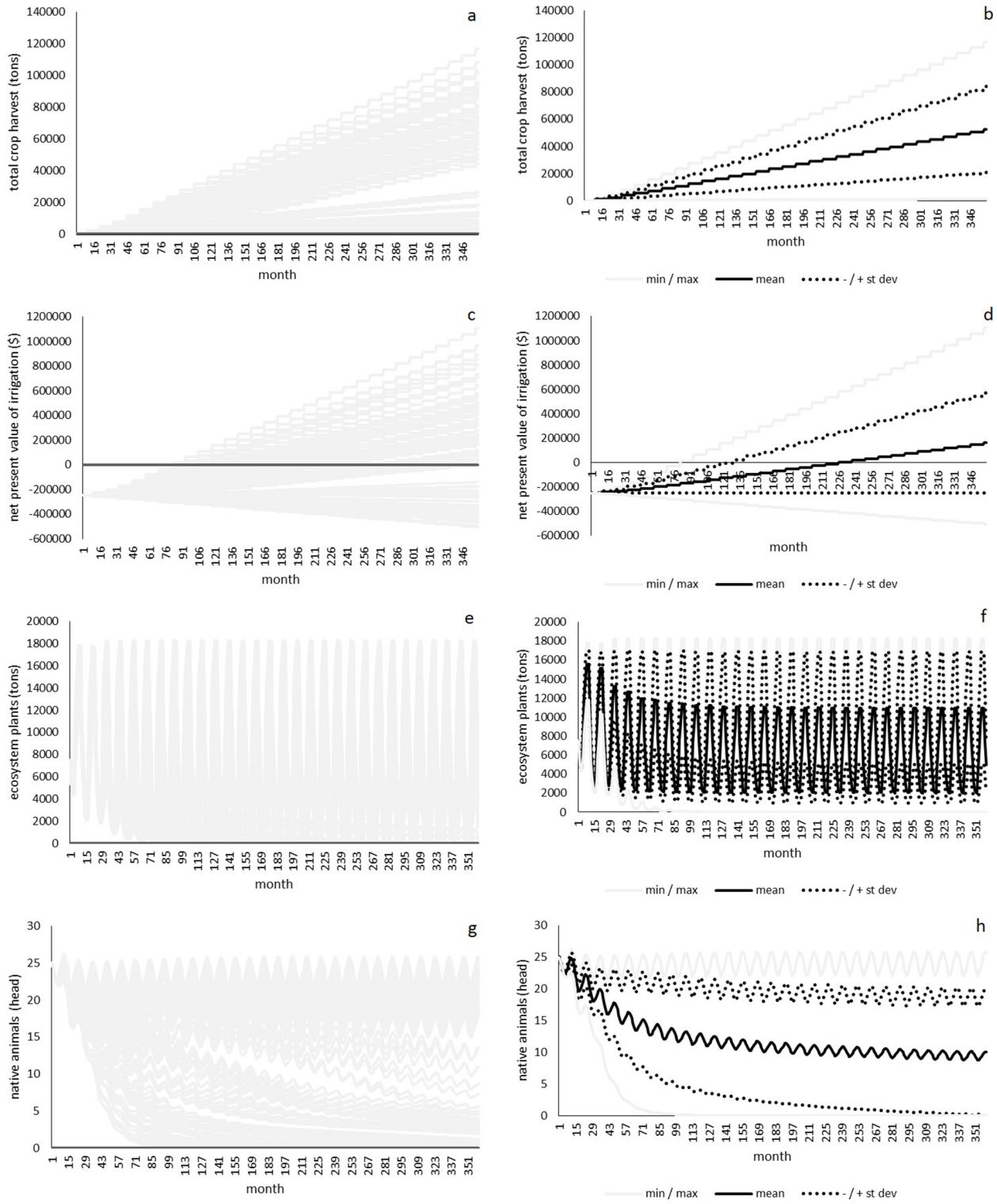

Fig. 10. Results of sensitivity analysis given 100 simulations of alternative parameter values for the eleven inputs: all results of total crop harvest (panel a) and the mean, minimum, maximum, and standard deviations in crop harvest (panel b); all results of net present value of irrigation (panel c) and its mean, minimum, maximum, and standard deviation (panel d); all results of ecosystem plants (panel e), and the ecosystem plants mean, minimum, maximum, and standard deviation (panel f); and all results for native animals (panel g) and the mean, minimum, maximum, and standard deviation of native animals (panel h).

divided by the final simulation time. Following Hekimoğlu and Barlas (2016), simulation runs were grouped by behavior pattern, pattern measures were estimated from each simulation run, values were standardized (Eq. (3)), and then input parameters were regressed to behavior pattern measures for each behavior pattern.

Regression equations using the eleven parameters in Table 5 were built to examine the effects that each had on the key output behavior pattern measure (Table 6). Regression results indicate the most significant parameters that lead to a particular behavior pattern. The regression equations for the two most common behavior patterns (goal-seeking decline with growth to equilibrium, $n = 39$; goal-seeking collapse, $n = 43$) were:

Table 5
Ending correlation coefficients (CC) between each input parameter used during sensitivity analysis with key stocks in the system, crop harvest, net present value (NPV) of irrigation, ecosystem plants, and native animals. Positive CC values indicate positive polarity (e.g., increasing base water diversions increases total crop harvest) while negative CC values indicate negative polarity (e.g., increasing base water diversions decreases ecosystem plants).

Input parameter	correlation coefficient with system variable of interest			
	total crop harvest	NPV of irrigation	ecosystem plants	native animals
base annual infrastructure costs	−0.236	−0.329	0.143	0.229
base crop planting density	0.266	0.156	−0.163	−0.285
base planting cost	−0.084	−0.122	0.051	0.003
base water diversions	0.578	0.516	−0.763	−0.698
decomposition rate	−0.010	0.030	−0.103	−0.115
discount rate	0.024	−0.152	0.098	−0.021
feed resource supplement	−0.059	−0.079	0.117	0.096
feed resource supplement cost	0.092	0.060	−0.105	−0.044
irrigation efficiency	0.580	0.525	−0.401	−0.317
price per ton-base crop	0.274	0.470	−0.165	−0.292
water consumption per ton of crop	0.120	0.098	−0.033	−0.013

Equilibrium level = 0.14base annual infrastructure costs – 0.36base crop planting density + 0.02base planting cost – 0.79base water diversions – 0.17decomposition rate + 0.03discount rate + 0.07feed resource supplement – 0.11feed resource supplement costs – 0.47irrigation efficiency – 0.21price per ton-base crop – 0.0003water consumption per ton of crop

Time of collapse = –0.37base annual infrastructure costs + 0.31base crop planting density + 0.19base planting cost + 0.40base water diversions – 0.10decomposition rate + 0.35discount rate + 0.11feed resource supplement + 0.06feed resource supplement costs + 0.02irrigation efficiency + 0.20price per ton-base crop – 0.20water consumption per ton of crop

Not shown in equation form are the results for inflection point and inflection level (used for behavior pattern 2 described above) since no significant parameters were identified, or for slope (used for pattern 4 described above) due to inadequate sample size. It was not surprising that no significant parameters were identified in the case of behavior pattern 2, which is the same generic behavior pattern as pattern 4 (goal-seeking collapse). The most significant parameter in goal-seeking collapse, *base water diversion rate*, also had the lowest p-values for both behavior pattern measures for pattern 2. Therefore, we may infer that the primary behavior mode expressed by the model is goal-seeking decline (n = 12 pattern 2 + 43 pattern 3 = 55 total goal-seeking decline) but that significant collapse is not induced unless the base water diversion rate reaches a high enough threshold.

In the equilibrium case, *base crop planting density, base water diversions, irrigation efficiency*, and *price per ton* all significantly influenced the equilibrium population size the native animal population was able to achieve. Importantly, all significant factors possessed negative polarities. For example, greater planting density, water diversions, irrigation efficiency, or crop price lead to lower equilibrium levels in native animals. This logically follows due to the trade-off between the agricultural system productivity and ecosystem productivity given feedback processes influencing the allocation of resources. The other major behavior pattern was goal-seeking decline to a nonviable population, where the behavior pattern measure was the time to collapse. In this case, the only significant factor was *base water diversion*, while other important variables were *annual infrastructure costs* and *discount rate*. Insignificance of other factors indicate less importance in creating the observed behavior pattern.

Results from both the statistical screening and behavior mode sensitivity analysis are summarized in Table 7. Key input parameters are ranked based on their influence and polarity on either maintaining a higher equilibrium of native animals versus the time to collapse in native animals in the goal-seeking collapse case.

4.4. Comparative analysis of alternative assumptions, decision rules, or policies (counterfactuals or what-ifs?)

Whereas the tests to this point have dealt with examining robustness and developing depth of system understanding (both of which primarily reside in model development and evaluation stages of the modeling process), the remaining tests focus on effectively using a model to generate insights needed for constructing and advocating for policy or strategy recommendations. Generating model insights via alternative assumptions, decision-rules, or policies can take the form of counterfactual trajectory analyses, boundary-adequacy tests on model behavior and policy recommendations, or intervention studies, each of which can be formulated using “what-if” experiments. These experiments provide opportunities to examine how emergent pressures arise or can be mitigated for, whether or not the effects of additional model structure in biophysical, decision-making, spatial components, or the interactions among them, changes the final model results and management recommendations, and to estimate the risk associated with adopting particularly new decisions or policies (Forrester and Senge 1980). We illustrate three such tests by posing specific what-if questions aimed at generating management insights via counterfactual trajectories, boundary-adequacy, and intervention thresholds.

4.4.1. Counterfactual trajectories

Counterfactual trajectories involve altering the basic model

Table 6
Regression results for the equilibrium level of goal-seeking decline with growth to equilibrium pattern, inflection point and level for goal-seeking decline behavior pattern, and time to collapse for goal-seeking decline pattern to a nonviable population size.

Input parameter	equilibrium level (n = 39)			inflection point (n = 12)			inflection level (n = 12)			time of collapse (n = 43)		
	stand. co.[1]	t-statistic	p-value	stand. co	t-statistic	p-value	stand. co	t-statistic	p-value	stand. co	t-statistic	p-value
base annual infrastructure costs	0.14	1.64	0.11	1.30	0.59	0.66	−0.32	−0.39	0.76	−0.37	−1.67	0.11
base crop planting density	−0.36	−4.01	**0.00**	2.10	1.27	0.42	−1.21	−1.96	0.30	0.31	1.42	0.16
base planting cost	0.02	0.30	0.77	0.19	0.16	0.90	0.15	0.35	0.79	0.19	0.79	0.43
base water diversions	−0.79	−10.44	**0.00**	9.39	1.31	0.41	−5.52	−2.07	0.29	0.40	2.08	**0.05**
decomposition rate	−0.17	−1.95	**0.06**	1.06	0.61	0.65	−0.72	−1.11	0.47	−0.10	−0.48	0.63
discount rate	0.03	0.40	0.70	0.74	1.21	0.44	−0.30	−1.33	0.41	0.35	1.75	**0.09**
feed resource supplement	0.07	0.81	0.43	1.03	1.09	0.47	−0.56	−1.60	0.36	0.11	0.46	0.65
feed resource supplement cost	−0.11	−1.29	0.21	−0.72	−0.82	0.56	0.45	1.38	0.40	0.06	0.28	0.78
irrigation efficiency	−0.47	−5.64	**0.00**	5.10	1.54	0.37	−2.88	−2.33	0.26	0.02	0.10	0.92
price per ton-base crop	−0.21	−2.74	**0.01**	1.21	1.07	0.48	−0.55	−1.32	0.41	0.20	0.91	0.37
water consumption per ton of crop	0.00	0.00	1.00	0.25	0.27	0.83	−0.21	−0.61	0.65	−0.20	−0.92	0.37

[1] standardized coefficient value.

Table 7
Summary of statistical screening and behavior mode sensitivity analyses indicating relative ranking of impact and polarity that input parameters have on native animal population.

Ranking	Equilibrium level of native animals	Time of collapse in native animals
1	base annual infrastructure costs (+)	base water diversions (-)
2	feed resource supplement (+)	irrigation efficiency (-)
3	base planting cost (+)	price per ton-base crop (-)
4	water consumption per ton of crop (-)	base crop planting density (-)
5	discount rate (+)	decomposition rate (-)
6	feed resource supplement cost (+)	feed resource supplement cost (-)
7	decomposition rate (-)	discount rate (-)
8	base crop planting density (-)	water consumption per ton of crop (-)
9	price per ton-base crop (-)	base planting cost (+)
10	irrigation efficiency (-)	feed resource supplement (+)
11	base water diversions (-)	base annual infrastructure costs (+)

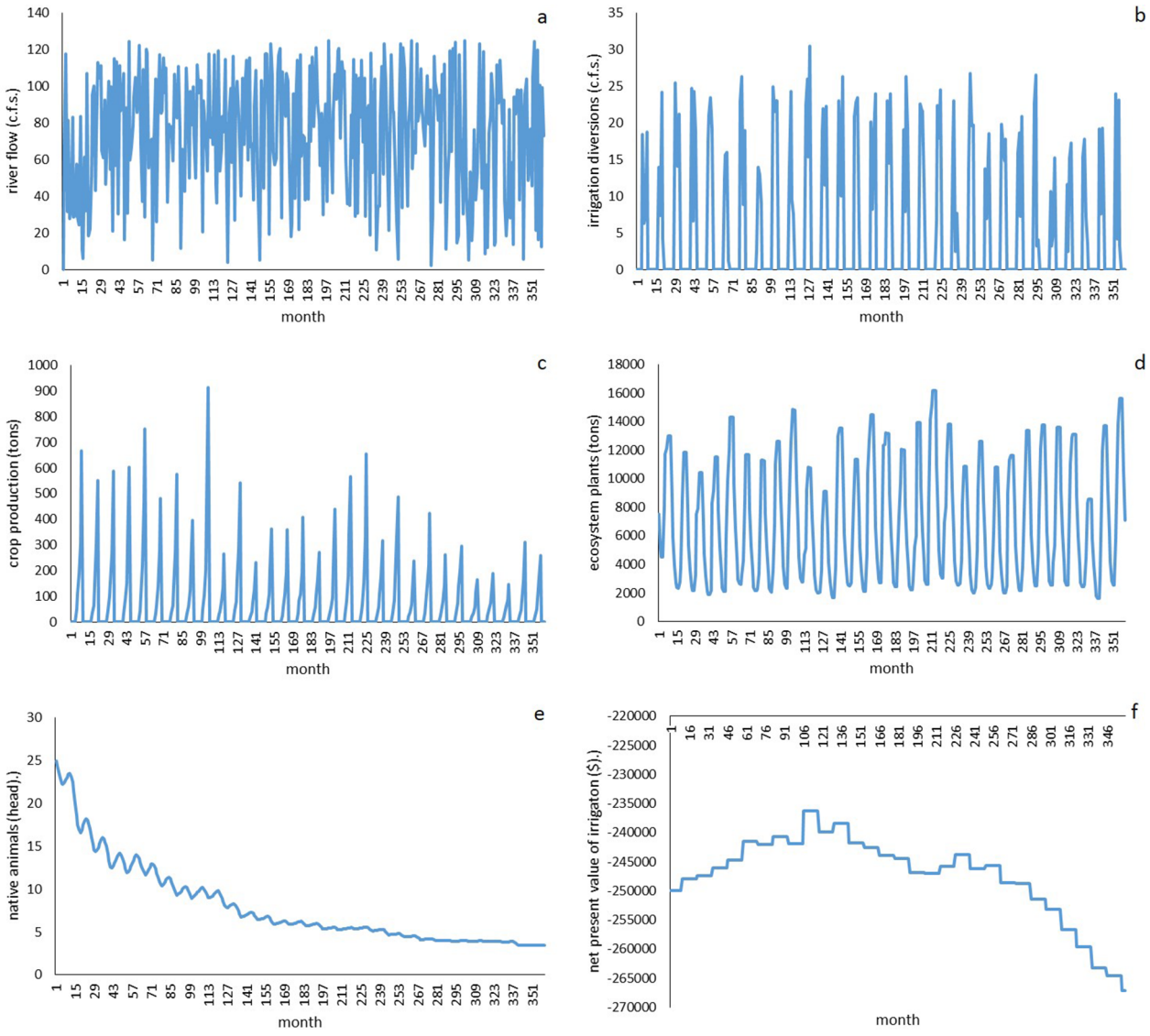

Fig. 11. Results of applying a counterfactual trajectory in river flow (panel a) to irrigation diversions (panel b), crop production (panel c), ecosystem plants (panel d), native animals (panel e), and net present value of irrigation (panel f).

assumptions or conditions in ways that are either known to be wrong, haven't been observed in the historical record, or were not assumed important enough during model development to be included in the model structure. A counterfactual trajectory is therefore a quantifiable and rigorous “thought experiment”. A key assumption of the model presented here is that there has been and will be a consistent, reliable surface water source based on existing information (calibrated to a long-run mean of 100 c.f.s.). A reasonable hypothesis would be that the insights generated by the model would be significantly different if a counterfactual condition regarding incoming surface water flows were used. Therefore, we posed the following what-if question: What if the river flow assumptions (100 c.f.s. across the time-horizon of simulation)

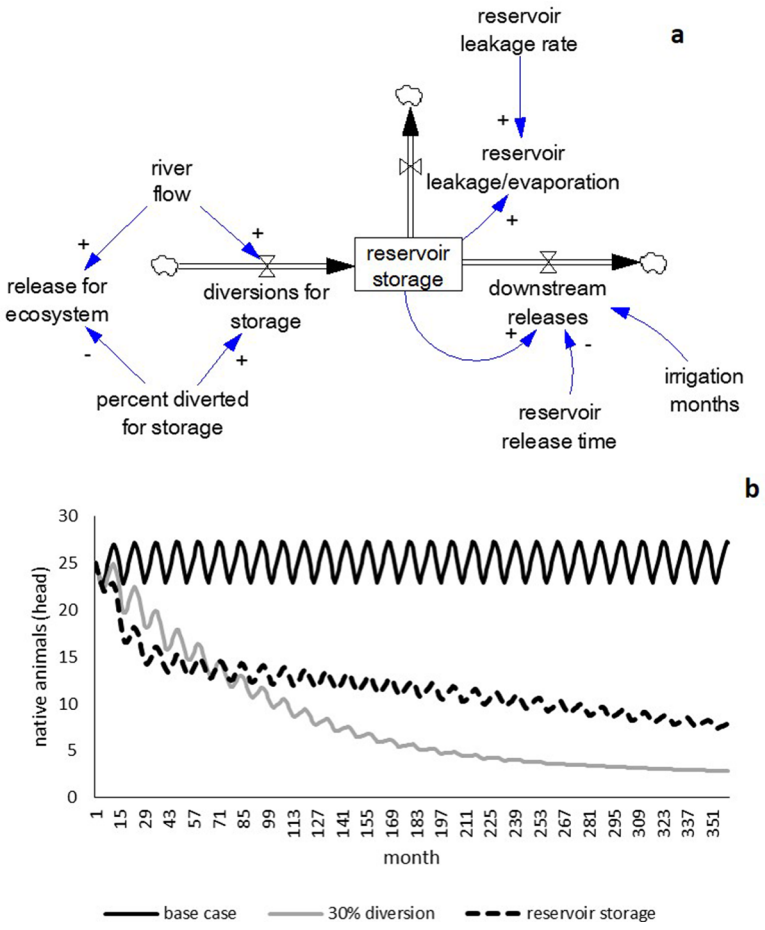

Fig. 12. Additional model structure representing reservoir storage and release added for boundary adequacy testing (panel a) and results of boundary adequacy (behavior) test illustrating changes in the native animal population under the base case and surface water diversion scenarios (30% diversion) to native animals given the expanded model boundary to include reservoir storage and release (panel b).

are incorrect? An alternative assumption is that river flow is highly variable due to climate and watershed characteristics, varying from zero (i.e., no-flow is the worst drought years) to 125 c.f.s (representing the occasional wet years of exceptional precipitation).

To test this counterfactual assumption in the model, minimum (zero c.f.s.), maximum (125 c.f.s.), and standard deviations (75 c.f.s.) in river flow rate were added to the model as auxiliary variable inputs to river flow. River flow, *rf* (recall as an input to Eq. (4)) then becomes a random function,

$$rf \sim N(\mu, \sigma^2) \quad (10)$$

where μ is the mean river flow (100 c.f.s.) and σ is the expected standard deviation in flow rate.[11]

Using this alternative assumption, we generate a dynamic rather than static trajectory in river flow (Fig. 11a). Due to the river flow variability, irrigation diversion rates during the growing season are less reliable (Fig. 11b), leading to a decline in crop production (Fig. 11c). This is due to the economic response in the agricultural sector, which alters cropping intensity based on changes in profitability (loop R1 in Fig. 2). The more profitable the agricultural sector, the greater investment in agriculture leading to greater crop intensity; the less profitable, lesser investment and intensity. This feedback contributes to the gradual decline in crop production as the intensity of production lessens over time. Ecosystem plants responded annually changes in flow (Fig. 11d), but because of habitat loss during droughts and in the most sensitive parts of the year the native animal population declines (Fig. 11e). Finally, the NPV of irrigation never approached breakeven-since crop production declines, revenues were not able to recoup costs of irrigation (Fig. 11f).

What management or policy recommendations would be altered given this counterfactual assumption in river flow? First, altering flows to include both wet and dry year cycles does not alleviate the pressure to the native animal population, as any benefit of improved habitat and per capita forage availability experienced during wet years is not enough to overcome the losses in dry years. Therefore, the management and policy concerns for the native population remains regardless of the flow assumptions compared here. However, new management pressures arise in the agricultural sector given the collapse in crop production and economic failure of investing in the irrigation system. If, given the alternative river flow trajectory, collapse in the agricultural sector is expected, policy-makers would be faced with the trade-off of either not approving the investment (if being analyzed prospectively), halting crop production before the irrigation investment NPV worsens after the first 10 years (if being analyzed retrospectively), or creating support mechanisms for the agricultural sector such that it is less susceptible to reductions in river flow (e.g., insurance that offsets annual losses; improvements in irrigation efficiency).

The first two responses would lead to improvements in native animal population since irrigation diversion would cease but would ensure a non-viable agricultural sector. In the third response, the decline in native animal population would only accelerate, since irrigation diversions would not be as responsive to changes in river flow (under an insurance scheme) or because of the reduction in return flows to the refuge (under improved irrigation efficiency). Therefore, given these counterfactual river flow assumptions, management and policy

[11] In the Vensim modeling environment, the built in function RANDOM NORMAL can be used, such as RANDOM NORMAL(river flow min, river flow max, river flow mean, river flow standard deviation, river flow seed value). Including the seed value provides means for adequate comparisons across simulations because the seed provides unique random sequence of values for each unique seed value.

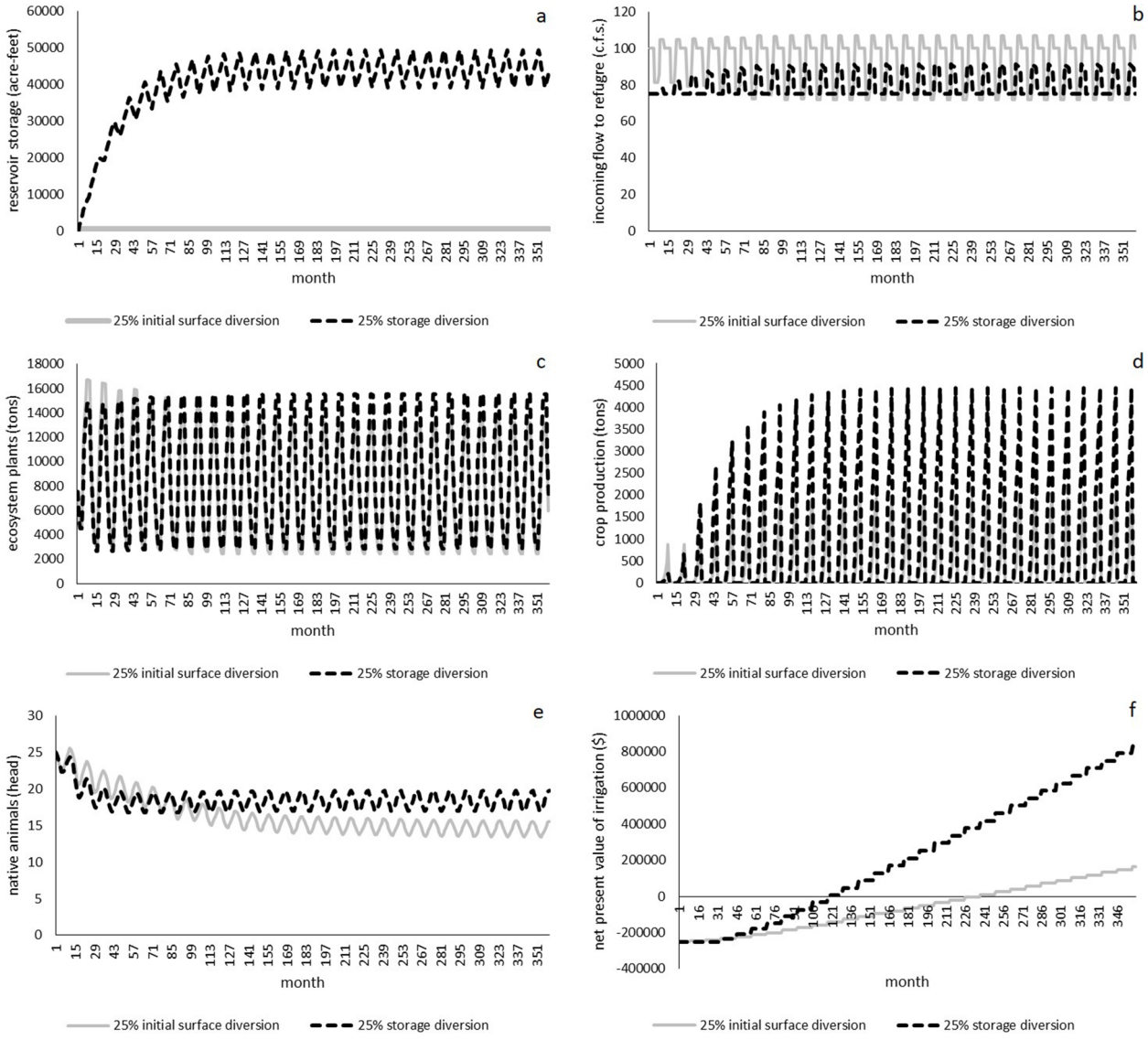

Fig. 13. Boundary adequacy (policy) test for optimal water diversion rates with and without reservoir storage and the resulting behavior patterns in reservoir storage (panel a), incoming flow to the refuge (panel b), ecosystem plants (panel c), crop production (panel d), native animals (panel e), and net present value of irrigation (panel f).

pressures arising due to agricultural collapse would erode conservation-oriented efforts to balance irrigation economics and the wildlife refuge. Under the original, static river flow assumption, reductions in base water diversion rate and irrigation efficiency were pinpointed as the most significant factors conducive of sustaining a stable native animal population (Table 4). Counterfactual trajectories such as these aid in examining the trade-offs and consequences of alternative assumptions (Appendix Section 3.4. provides an example of a failed counterfactual tests).

4.4.2. *Boundary-adequacy testing*

Boundary adequacy testing can be used in the context of model structure (is the model boundary appropriate given the model purpose?), behavior (can new model structure be conceptualized that significantly alters its behavior?), or policy (how does modifying the model boundary alter policy recommendations?; Forrester and Senge, 1980). Here we focus on model boundary tests in the context of behavior and policy recommendations. The current model boundary (Fig. 2) encompasses the wildlife refuge that is habitat for the ecosystem plants and native animals, the surrounding cropland, and the river flow that supplies surface water diversions for irrigation. A relevant boundary-adequacy test would be expansion of the model boundary to include a new state (stock) variable capable of expressing its own unique dynamics. After adding the new structure, we can then examine how the expanded model may influence the problem of interest. In this case, we expand the model boundary by asking: what-if the watershed constructed a reservoir system for irrigation deliveries that would help regulate river flow and avoid the effects caused by seasonal river flow diversions for irrigation? The hypothesis here would be that the reservoir would help redistribute the water supply from the non-growing season to the growing season thereby reducing the impact of irrigation to the refuge. To complete this test, additional model structure was added representing reservoir storage and release (Fig. 12a).

In the added structure, a certain percentage of river flow (30%) is diverted for storage with the remainder immediately released for the ecosystem (70%). The diversions enter a reservoir storage, which incurs losses due to natural leakage and evaporation (0.001%) and through downstream releases. Releases downstream are a function of the reservoir release time (capturing average residence time of water in storage, ≈12 months) plus the water released for irrigation during the crop growing season (i.e., no water for irrigation is released during the

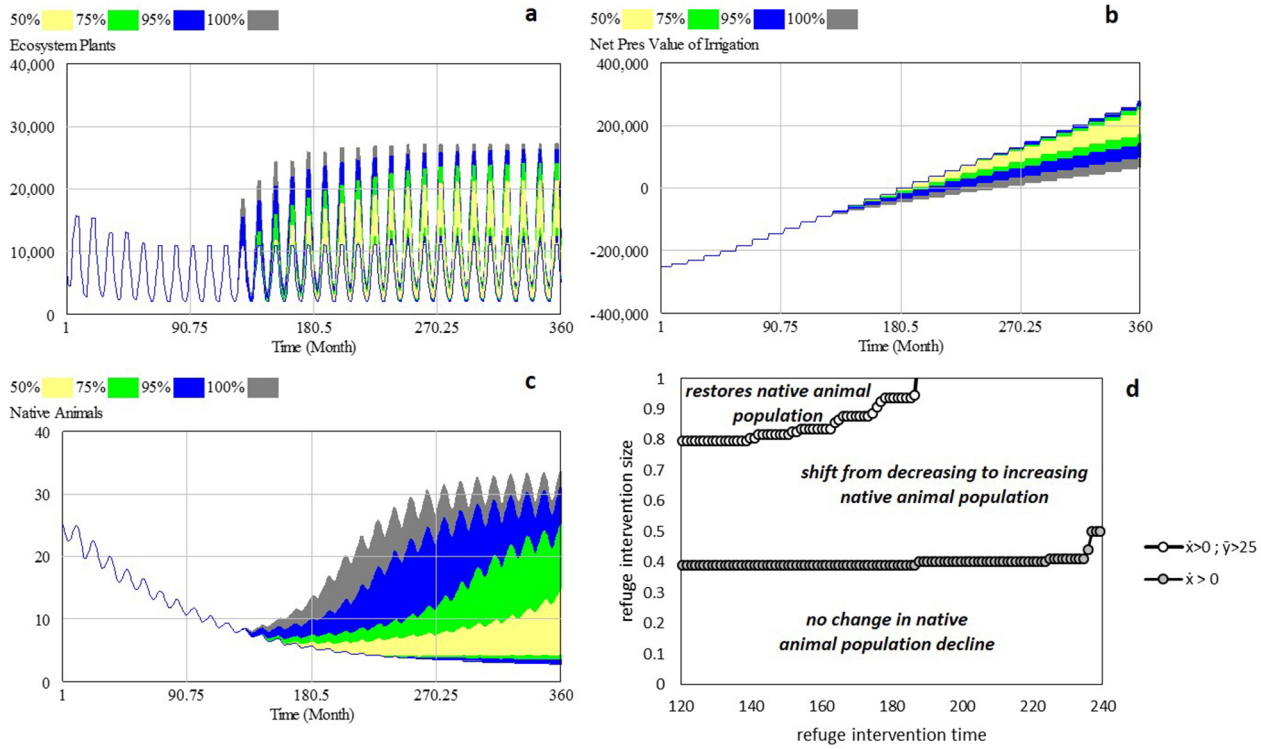

Fig. 14. Intervention threshold analysis results given 10,000 simulations of unique combinations of intervention size and time: percentile intervals for ecosystem plants (panel a), net present value of irrigation (panel b), and native animals (panel c); intervention threshold graph (panel d) illustrates the minimal refuge intervention size at each particular intervention time needed to shift the native population from decline to growth or decline to growth and restoration of original population size.

non-growing season). Since no new agronomic assumptions are included in this scenario, once water for irrigation is delivered it remains subject to the same irrigation efficiencies and return flow rates as the base case. By storing water during the non-growing season and releasing it throughout the year, we hypothesize that a different behavior pattern will arise in the native animal population. The resulting behavior patterns given the expanded model boundary are then compared to the behaviors prior to model boundary expansion. Adding the reservoir storage and release structure did improve the native animal population size relative to 30% direct surface water diversions for irrigation, but did not change the overall behavior pattern (Fig. 12b). This is a strong test of the model boundary that, due to the resulting behavior in native animals, strengthens confidence in the original boundary. If the test resulted in a different behavior-mode in native animals, reexamining the original model structure and its links to the native animal population would be warranted.

An additional boundary experiment includes examining alternative policy recommendations that would arise given the expanded model boundary and structure representing reservoir storage. First, we identify the management recommendation concerning irrigation diversions under the original model boundary and given the policy-goal is to achieve a stable native animal population in equilibrium that still allows for profitable crop production. We then identify the management recommendation under the expanded model boundary and compare that to the original recommendation. In order to identify these points, the base water diversion rate used for determining irrigation was manipulated by hand until the policy-goal conditions were reached.

Under the original model boundary, the base water diversion rate found to balance native animals without jeopardizing crop production and profitability was 25% of river flows. Without any reservoir storage potential (Fig. 13a), incoming river flow to the refuge is anchored to the total river flow, 100 c.f.s., in the non-growing season, but is subject to large declines during the growing season (to as low as 70 c.f.s) when the refuge needs water the most (Fig. 13b). Due to return flows, the incoming flow to the refuge peaks over 100 c.f.s., but the marginal benefit of the additional flow is negligible since it occurs after the primary growing season. Ecosystem plants, with the reduction in incoming river flows during the growing season, declines to a peak of 12,000 tons (Fig. 13c), while crop production peaks at 2,000 tons (Fig. 13d). The goal to maintain a viable native animal population and a profitable agricultural system is achieved, with mean animals ending at 15 head (Fig. 13e) and NPV of irrigation ending above $200,000 (Fig. 13f).

Under the expanded model boundary, the storage water diversion rate (i.e., the percentage of river flow diverted to storage) found to balance native animals without jeopardizing crop production and profitability was also 25%. Given that diversion rate, reservoir storage would need to be capable of storing between 40,000 and 50,000 acre-feet of water (Fig. 13a). Because the river flow is being redistributed with storage, the incoming flow to the refuge was anchored at 75 rather than 100 c.f.s. (Fig. 13b). However, the refuge receives more water during the growing season, up to 90 c.f.s., due to the releases from storage combined with return flows from the irrigation system. This leads to a more stable ecosystem plant community (Fig. 13c) and greater irrigation levels supportive of almost double crop production (Fig. 13d). With a more stable plant community, native animals reach equilibrium near 20 head (+33% compared to the original recommendation above). The NPV of irrigation reached over $800,000 (or a 400% increase over the original recommendation) due to increased crop productivity. As observed here, the expanded model boundary would lead to significantly different management recommendations (Appendix Section 3.4. provides examples of failed boundary-adequacy tests one might encounter).

4.4.3. *Determination of intervention thresholds*

Finally, we estimate the intervention thresholds (i.e., the minimum intervention size and intervention time that results in the desired

behavior change) for the native species population. As shown in Fig. 3, the atomic behavior pattern exhibited by the native animal population given irrigation diversions is goal-seeking decline. The desired post-intervention behavior would be s-shaped growth, which would reflect a population that increases up to a new equilibrium level equal to the baseline population prior to irrigation diversion (potential behavior shown Appendix Fig. A1). Assume that construction of the reservoir system, described in the previous section, would not be approved due to other social, environmental, and economic concerns (e.g., rural community relocation, fish habitat connectivity, uncertain NPV due to regulatory and litigation costs). Other possible interventions include reintroducing new animals once the population declines below a population threshold determined by the policy-makers or importing feed resources as supplement to offset losses in ecosystem plants.

Both of these interventions are unfortunately low leverage. We can infer from the previous experiments the chain of causality driving native animal population declines: the feedback process between ecosystem plants (regulated by river flow) and native animals via per capita forage availability. Declines in native animal population are a symptom of the problem which, at the structural level, arises where river flow supplies the refuge. Reintroducing animals or supplying feed resource supplementation are symptomatic-solutions which may work only in the short-term and at a very high cost to the system (Appendix Section 3.4. provides the simulation evidence for these). A more fundamental solution would be address the problem at the refuge level. For example, what if rather than focusing on symptomatic solutions of the native animal population (via policy-interventions that "prop up" the population), investment is made in expanding the wildlife refuge? Under such a scenario, effort may be made to improve the surrounding habitat for the population and expand the wildlife refuge, thereby increasing the per capita forage availability.

In order to implement this test, a refuge expansion intervention function is added to the model (similar to the step equation in Eq. (6)) where a specified *refuge expansion rate* (zero to 100%) and *refuge expansion time* are determined. The refuge expansion intervention applies to the stock of ecosystem plants based on the *refuge expansion rate* (e.g., a 100% *refuge expansion rate* would double the size of the current refuge). Expanding the suitable habitat would create additional land costs, but with greater forage availability for native animals, we would hypothesize that the behavior pattern of native animals is shifted from a decline-oriented behavior to one of s-shaped growth and stabilization.

To test this hypothesis and identify the minimum intervention needed for returning the native animal population back to its initial state (mean 25 animals), we apply a Latin grid experimental design (similar to the sensitivity tests in Section 4.3.3) with the input parameters being *refuge expansion rate* (ranging from zero, no expansion, to one, doubling the refuge size, in increments of 0.01) and *refuge expansion time* (ranging from 120 months to 240 months in increments of 1.2 months). This Latin grid experimental design (100×100) resulted in 10,000 simulations, one for every unique combination of *refuge expansion rate* and *refuge expansion time*.[12] The Latin grid design is preferred over Monte Carlo simulation in this case because of the explicit interest in identifying the specific *intervention size* and *time* that results in the desired behavior pattern. Use of Monte Carlo simulation risks duplicating certain input values (or at least very near combinations of values) and does not guarantee that all possible combinations of input values will be sampled. A Latin grid design ensures all possible combinations of input values are sampled and, although computationally intensive, is still more efficient than the required sample size to achieve the necessary input combinations using Monte Carlo simulation.

Following Walrave (2016) an indicator value was specified to determine if the desired behavior shift was achieved. In this case, the indicator variable used was the moving average of native animals and its first derivative.[13] Mean native animals was chosen as the behavior pattern measure given that the behavior of native animals inherently oscillates (due to annual reproduction and mortality dynamics) and because the oscillation can grow or decline as ecosystem conditions change such as it does with the onset of irrigation (i.e., the population does not oscillate around a stable fixed point). Using the moving-average smooths out these oscillations, resulting in a behavior where the signs of the first and second derivatives are quite stable. Under goal-seeking decline the value of the first derivative of mean native animals is negative and under goal-seeking growth it is positive. Therefore, we monitor the number of sign changes in the first derivative of mean native animals from negative to positive. If a change in sign does occur, we then identify if the mean population size reaches its desired level prior to irrigation (25 animals).

The results of the intervention experiment are shown in Fig. 14, including the percentile intervals and intervention threshold graph. Percentile intervals display the percentage of simulations falling within a particular range at a given point in time. As expected, ecosystem plants increased as a result of refuge expansion (Fig. 14a), with over half of the intervention combinations resulting in peaks up to 21,000 tons. Due to the additional land costs involved with improving the surrounding habitat and expanding the refuge, the NPV of irrigation does decline, but even the most extreme outcomes result in positive NPV near $100,000 (Fig. 14b). Importantly, these two outcomes illustrate that the intervention strategy does expand the refuge and can do so without financially taxing the system to an unprofitable level.

The behavior pattern changes in the native animal population resulting from the intervention combination were much more dynamic relative to ecosystem plants and NPV of irrigation (Fig. 14c). Nearly half of the simulations made no shift in behavior pattern (40% of all simulations). Of the 60% of simulations that did create a behavior pattern shift from goal-seeking decline to s-shaped growth, only 9% achieved a mean native animal population of 25 animals.

Using the resulting simulation data, an interventions threshold graph was constructed to illustrate the combination of minimum *refuge intervention sizes* and *refuge intervention times* that would be required to achieve an s-shaped growth to equilibrium of at least 25 animals (Fig. 14d). In order to shift the behavior pattern from decline to growth, the minimum *refuge intervention size* was 39% in month 120, increasing up to 50% in month 240 (denoted as $\dot{x} > 0$). Any *refuge intervention size* below 39% would not reverse the native animal population decline. In order to restore the native animal population (denoted as $\dot{x} > 0$; $\bar{y} > 25$), the minimum *refuge intervention size* would have to be at least 79% at month 120, up to 95% by month 186. After month 186, the intervention size would have to be more than double the size of the current refuge (over 1 on Fig. 14d, outside the range of values performed in the experiment). The resulting graphs indicate the minimum combination of intervention inputs needed to create a shift the dominant feedback processes (i.e., tipping point) from goal-seeking decline (negative feedback) to s-shaped growth (positive feedback). An important note regarding intervention thresholds such as these is that the intervention combinations included in the experiment had dissimilar incubation times due to a fixed end point of the simulation at 360 months. Because of this, novel behavior characteristics may be

[12] In the Vensim modeling environment, these simulations were completed in under five minutes. With the grid selection based on refuge expansion rate, the 10 year or 120 month period for refuge intervention time was not evenly distributed. Therefore, the simulation results were matched to the nearest whole month.

[13] The natural behavior mode for native animals is oscillation, but with the onset of irrigation, the overall pattern becomes goal-seeking decline. However, the infinitesimal changes in native animals, even under goal-seeking decline, still express oscillations albeit with smaller and smaller amplitudes. Because the first derivative of oscillatory behavior modes is misleading, we use the first derivative of the moving average of the stock (Walrave 2017).

exhibited beyond month 360 that would not be observable with a fixed final simulation time, but could be included under a dynamic final simulation time.

5. Conclusion

Dynamic systems models are increasingly used by scientists, managers, and policy-makers due to the growing complexity and interdependency of problems that persist in ecologic and socio-economic systems. Coupled to this is the growing awareness that, when confronted with such complexity, our human intuition rarely properly infers the underlying dynamics driving decision-making and its outcomes. Formal mathematical models are therefore essential tools for improving our understanding and decision-making in the face of such complexity. The above experimental examples illustrate a number of key tests (with increasing degree of difficulty from novice to advanced skill levels) any investigator can and should perform in order to evaluate and test their particular model. Each of the experiments demonstrate several key lessons: 1) that model experiments help uncover unforeseen flaws or incorrect formulations, including flaws in our own mental models, since all models are based on our mental representations of a given problem or system; 2) that good model experiments provide a means to glean valuable insights about the structure and behavior of a model; and 3) model experiments enhance the confidence (validity) of the model for its intended use, especially after the iterative process of identifying errors or flaws in model structure or behavior, updating and improving our mental model, and then revising the formal simulation model in turn. The comprehensive suite of tests explored is not an exhaustive list of model testing procedures as there a number of other advanced methods for understanding uncertain parameter values, identifying dominant feedback structures, and testing alternative decision-making theories. Although insightful, such tests are beyond the skill-level of novice modelers and depending on the problem and objective of the model, the insights generated may not be important to the issue or efficient to attain given the required investment in time and resources.

Novices should be aware of what expert modelers will recognize from the experiments and discussion presented above – that model development and experimentation is an iterative process, often requiring numerous iterations of experimentation, analysis, and revision (to both mental and simulation models; Fig. 1). The experimental results and discussion in this paper are the final product of that iterative process. In early stages of development, a model will not perform adequately when exposed to the barrage of tests outlined above and novices will soon find errors or omissions in model structure that lead to implausible or unexpected behaviors that require explanation and correction. The model presented above was exposed to numerous rounds of revision and correction prior to "passing" the model behavior experiments. Then it was exposed to comprehensive sensitivity analyses. After adequately withstanding the sensitivity tests, the model moved forward to the more complex experiments dealing with alternative assumptions and structure. Novices should not be discouraged when their models fail any of these tests along the reiterative modeling process. Failing any one test can be expected and provides the needed feedback to the modeler about which model component needs to be improved prior to looking for strategy or policy insights to guide decision-making.

The increasing interest in and more frequent application of systems models by ecologists, agriculturalists, and natural resource managers is a positive indicator of the recognition of modeling as a valuable tool to better understand and manage the complex, dynamic systems that we operate in. Employing such models has the potential to increase our understanding of the many, poorly understood feedback processes that must be well-managed if such systems are to function as society desires. If researchers who begin or are currently using systems modeling approaches do not have at least a basic understanding of the highly iterative, systems analysis process or of the fundamental experimental tests that are required to build confidence that the resulting model is trustworthy to be used, the potential for making management and policy recommendations based on fundamentally flawed analyses (i.e., incorrect and unreliable model recommendations) is high. This paper provides a basic introduction and guide to experimental testing of a developed model in order to build confidence that the model is capable of providing robust insight into a given problem. By providing this guide, we hope that more modelers will be better prepared to build, evaluate, and test their models such that the resulting model-generated insights will be more capable of mitigating unintended consequences or improving desired functions of the complex ecological or natural resource systems we are tasked with managing.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

This work was partially supported by United States Department of Agriculture's Higher Education Challenge Grant No. 2018–70003–27664 for "Curriculum Development for Wicked Problem Solving" and the National Science Foundation's Center for Research Excellence in Science and Technology (CREST) Award No. 1914745, both of which the author is a Principal Investigator. The author also wishes to thank Dr. Luis Tedeschi, Dr. Barry Dunn, Mr. Michael Goodman, Mr. Corey Peck, and Dr. William Grant and for their helpful comments and suggestions pertaining to both general modeling and pedagogical issues prior to completion of the manuscript as well as two anonymous reviewers for their thoughtful and constructive comments which greatly improved the paper.

Supplementary materials

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.ecolmodel.2020.109246.

Appendix

1. *Understanding behavior modes*

Behavior modes expressed over time are indicators of the underlying feedback processes that interact to produce the observed behavior pattern of a particular variable of interest in a system. From a systems thinking perspective, these feedback processes are often simplified into single-loop positive (reinforcing) or negative (balancing) processes perceived to be the dominant parameters or feedback structure. Unfortunately, without a rigorous quantitative analysis, identification of the most influential parameters and feedback structure is spurious at best. Therefore, it is important to be able to differentiate between alternative behavior modes via their mathematical indicators (e.g., the first- and second derivatives; Fig. A1).

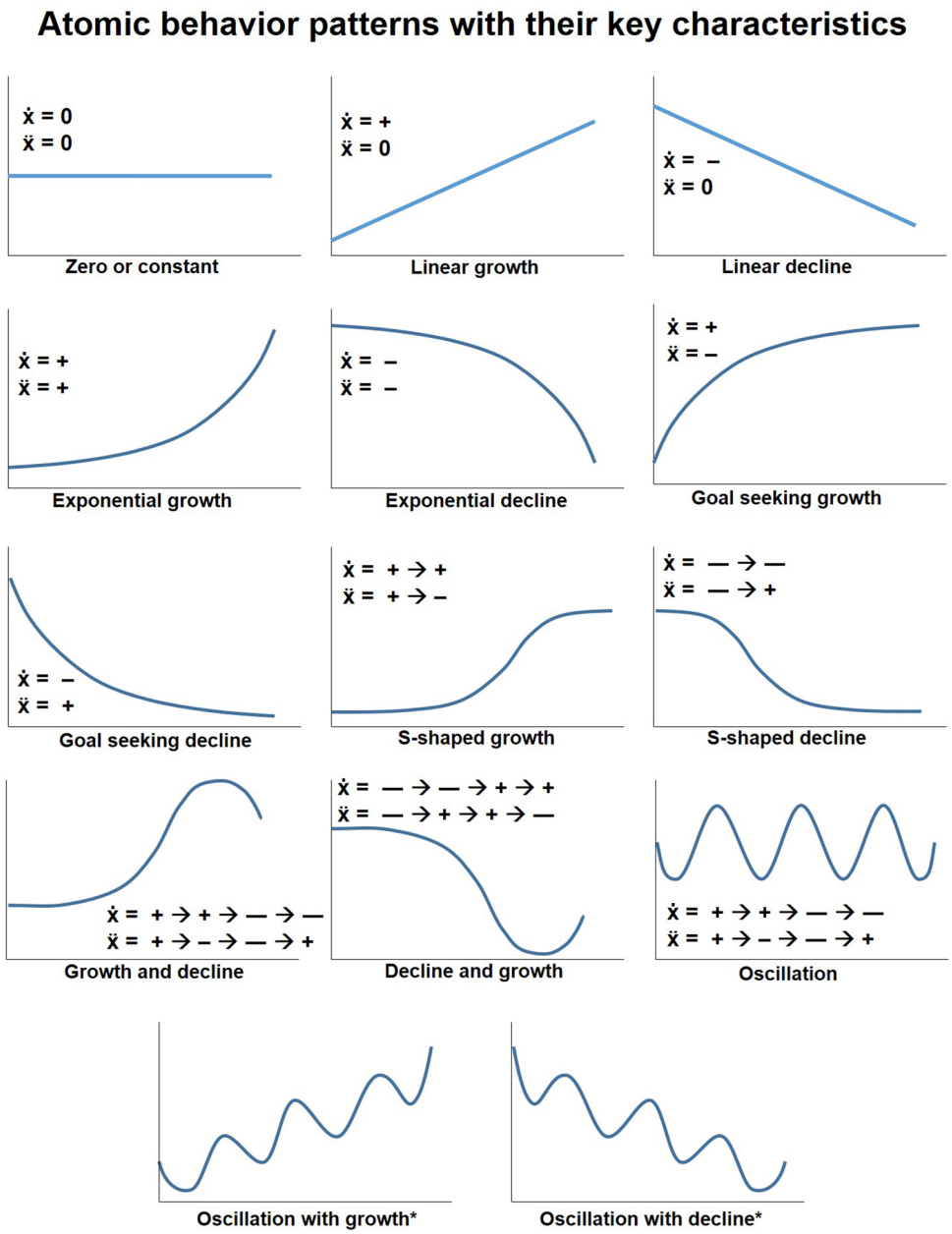

Fig. A1. Atomic behavior patterns and their associated key characteristics observed in the first (ẋ) and second (ẍ) derivative.

2. *Model documentation*

Included in the supplementary material are: a) a copy of the model formulated in Vensim modeling environment (Ventana Systems); b) Microsoft Excel file that includes all simulation results and is a template for analyzing the results of the experiments; and c) data file of the intervention thresholds results. Fig. A2 provides an expanded conceptual model of Fig. 2 in the paper for additional details regarding auxiliary variables included around the core feedback structure.

3. *Examples of common model revisions made during iterative model testing process*

One may observe that the model presented in the paper successfully passed all of the performed tests. As noted in the text, modeling is an iterative process, the presented results being the end of a longer series of experimentation and model revision. In order to illustrate how these techniques and tests work together in practice in an iterative model development, testing, and revision process, this appendix section provides examples of situations where the model (either mental or quantitative) failed a test and therefore required revision of the model. The revisions here are meant to illustrate common mistakes that many, particularly beginners, may encounter, not an exhaustive protocol for model error identification.

3.1. *Extreme conditions tests*

There are several common errors or mistakes one may find when the extreme conditions test is failed. Consider the example from section 4.3.1 where *river flow* is increased from 100 c.f.s. to 150 c.f.s., and the resulting behavior pattern observed in *ecosystem plants* and *native animals* is

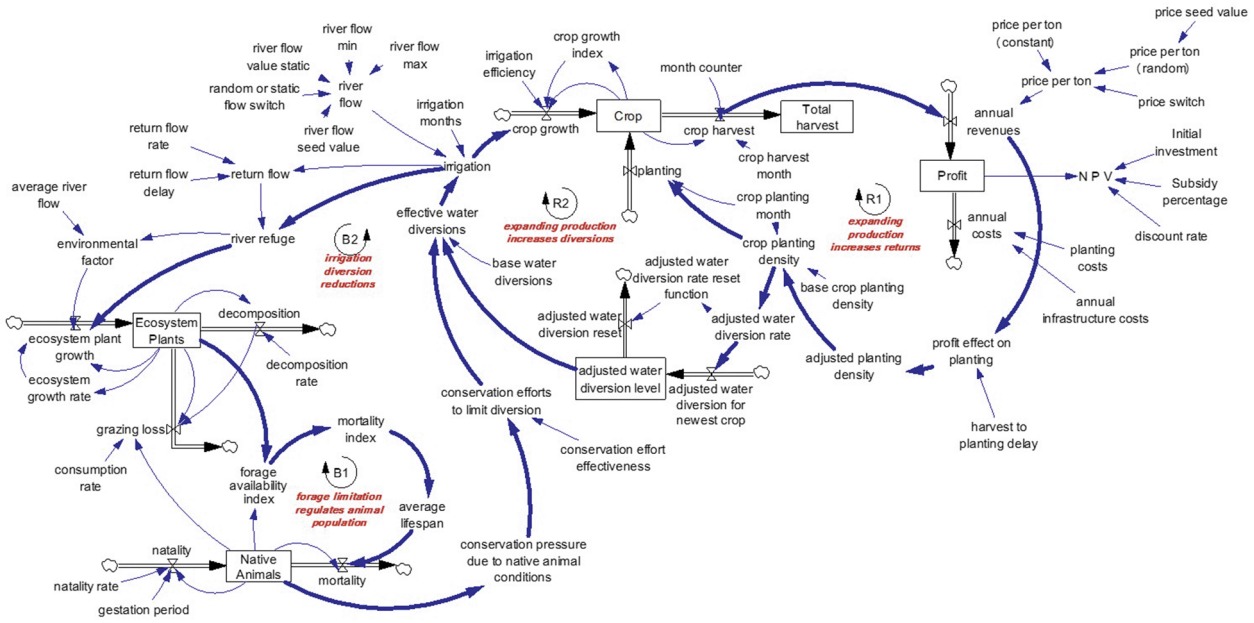

Fig. A2. Expanded stock-and-flow diagram of the irrigation-wildlife case-study model.

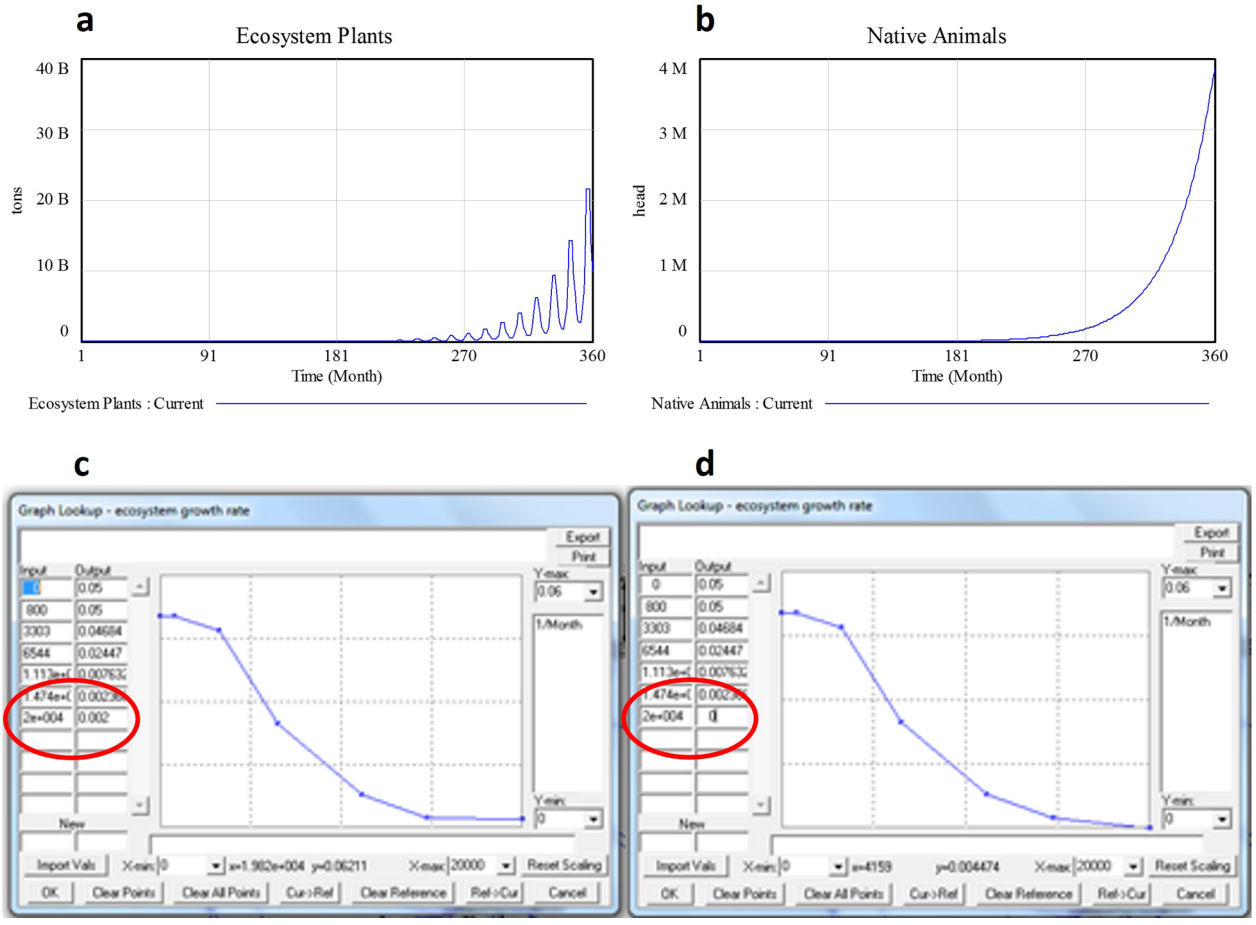

Fig. A3. Example of failed extreme conditions test due to open-ended bound on the ecosystem growth rate function (see Fig. A2).

runaway, exponential growth (Fig. A3a and b). In order to correct this, we may inspect the stock-flow and auxiliary structures around *ecosystem plants* and *native animals*. We may find that we originally parameterized the graphical function ecosystem growth rate with a minimum value of 0.002 at 20,000 tons of *ecosystem plants*. Because many dynamic modeling programs default to extrapolation of graphical functions at the end points for input values beyond the parameterized range, having a positive value for the growth rate at the estimated biophysical maximum value of 20,000 tons means that for any value of *ecosystem plants* above 20,000 tons defaults to a 0.002 growth rate per month (Fig. A3 panel c), leading to greater

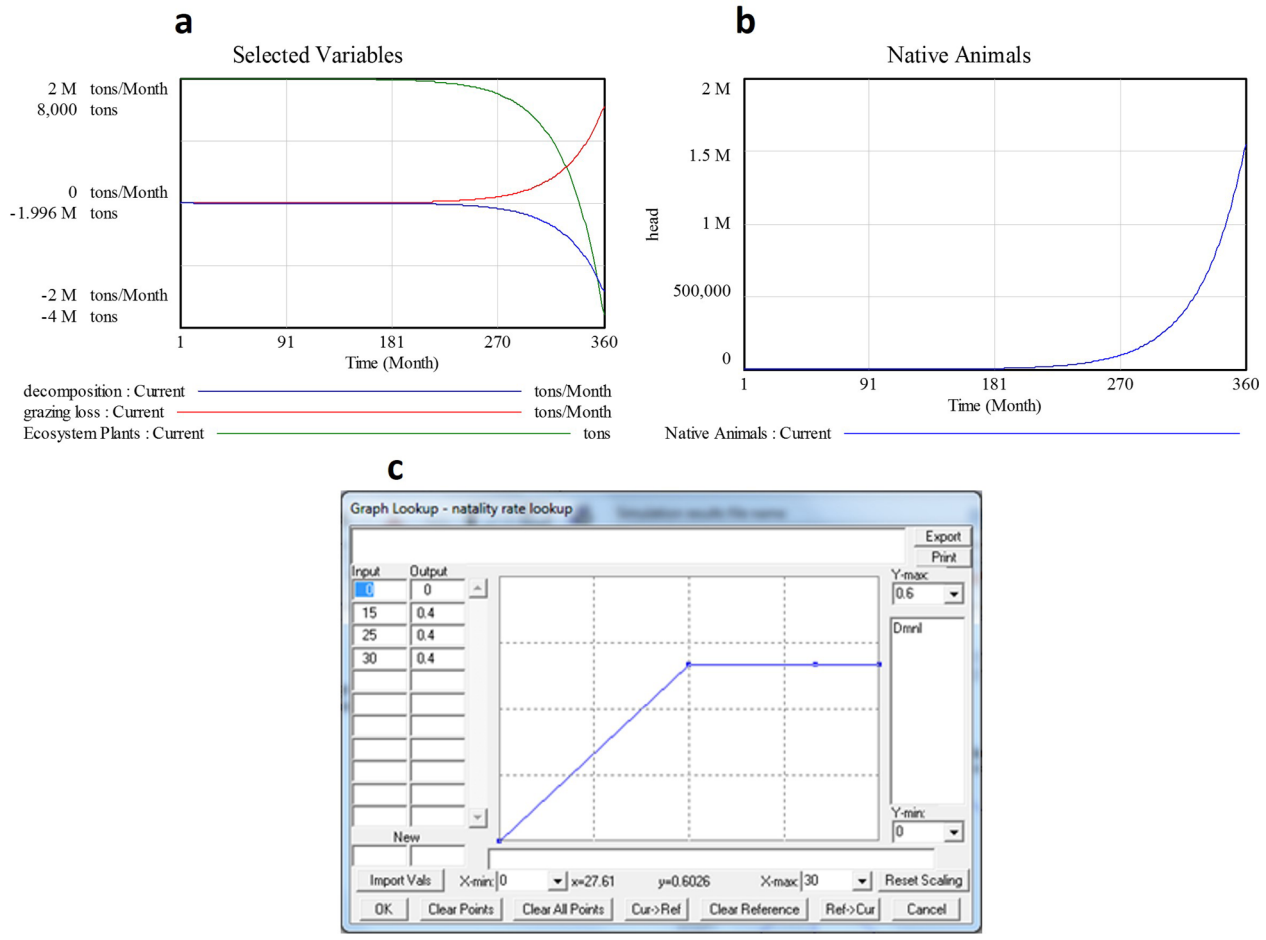

Fig. A4. Example of failed extreme conditions test due lack of first-order negative feedback processes on stock variables and static natality parameter value that necessarily lead to runaway exponential growth (panels a and b), with the inclusion of a dynamic natality rate (panel c).

ecosystem plants subject to the same 0.002 growth rate per month. This leads to the runaway growth behavior in Fig. A3 panel a. The *native animal* population, without any forage limitation, also grows in conjunction with *ecosystem plants* (Fig. A3 panel b). By adjusting the *ecosystem growth rate* to 0 at the biophysical maximum for the habitat in the refuge, 20,000 tons, means that any values of *ecosystem plants* above 20,000 tons will default to 0 for the growth rate (Fig. A3 panel d), and therefore *ecosystem plants* will cease to grow exponentially.

Suppose that change is made and we expect to see the problem corrected in the model, yet we still observe runaway exponential growth in *native animals* and runaway exponential decay in *ecosystem plants* (Fig. A4). We again examine the stock-flow and auxiliary variable structures around each variable. We find that the resulting behavior pattern is due to *grazing loss* to *ecosystem plants* increasing exponentially (Fig. A4 panel a) due to the fact that grazing loss is directly proportional to *native animals* (Fig. A4 panel b). In reality, we know that *ecosystem plants* and *native animals*, being a physical quantities, cannot take on negative values (in the case of *ecosystem plants*) or grow forever despite the fact that there are no resources to do so (in the case of *native animals*). Further checks of the model reveals that *native animals* was first parameterized with a static *natality* values, with *natality* possessing a greater value than the *mortality index*. When *natality* > *mortality*, the result is exponential growth in the population. In order to alleviate this issue, a dynamic *natality rate* was developed (Fig. A4 panel c). The assumption here would be that if the population declines below the expected equilibrium value of 25 animals, the *natality rate* also declines due to the greater energy requirements for searching for and finding reproductive mates. Correcting this should bring the dynamic *natality rate* into equilibrium with the *mortality rate* (a function of the *forage availability index*, Fig. A2). By doing so, *native animals* (and therefore grazing loss) will be constrained. In addition, we don't want *ecosystem plants* to take on negative values. Examining the outflows reveals that the model lacked first-order negative feedback controls to regulate the stock-flow dynamics to be physically conserved. In the first instance, grazing loss was simply a function of *consumption rate* (= *native animals* X *forage consumption per month*). With the revised model, grazing loss is regulated by *ecosystem plants* such that the grazing cannot exceed the available biomass in the ecosystem plant stock [*grazing loss* = MIN(*consumption rate, ecosystem plants – decomposition*)].

3.2. Step, pulse, and ramp functions

Examining how a model responds to various steps, pulses, or ramps is an valuable exercise in developing depth of system understanding, and like extreme conditions tests, helps identify places for model improvement when the simulation results fail to align with expected or observed behaviors in the real-world system.

First, consider a step change in planting costs (the most significant annual input cost to the irrigation system) from $50 to $100 dollars per ton (Fig. A5 panel a). Due to the increase in cost, we would expect profitability to decline, leading to feedback sequence of declines in subsequent planting and irrigation rates, crop production, and finally long-term profits. However, we see that annual costs decline into negative values (Fig. A5 panel b). Because annual costs are an outflow from cumulative profits (Fig. A2) and stock or state variables are typically difference equations

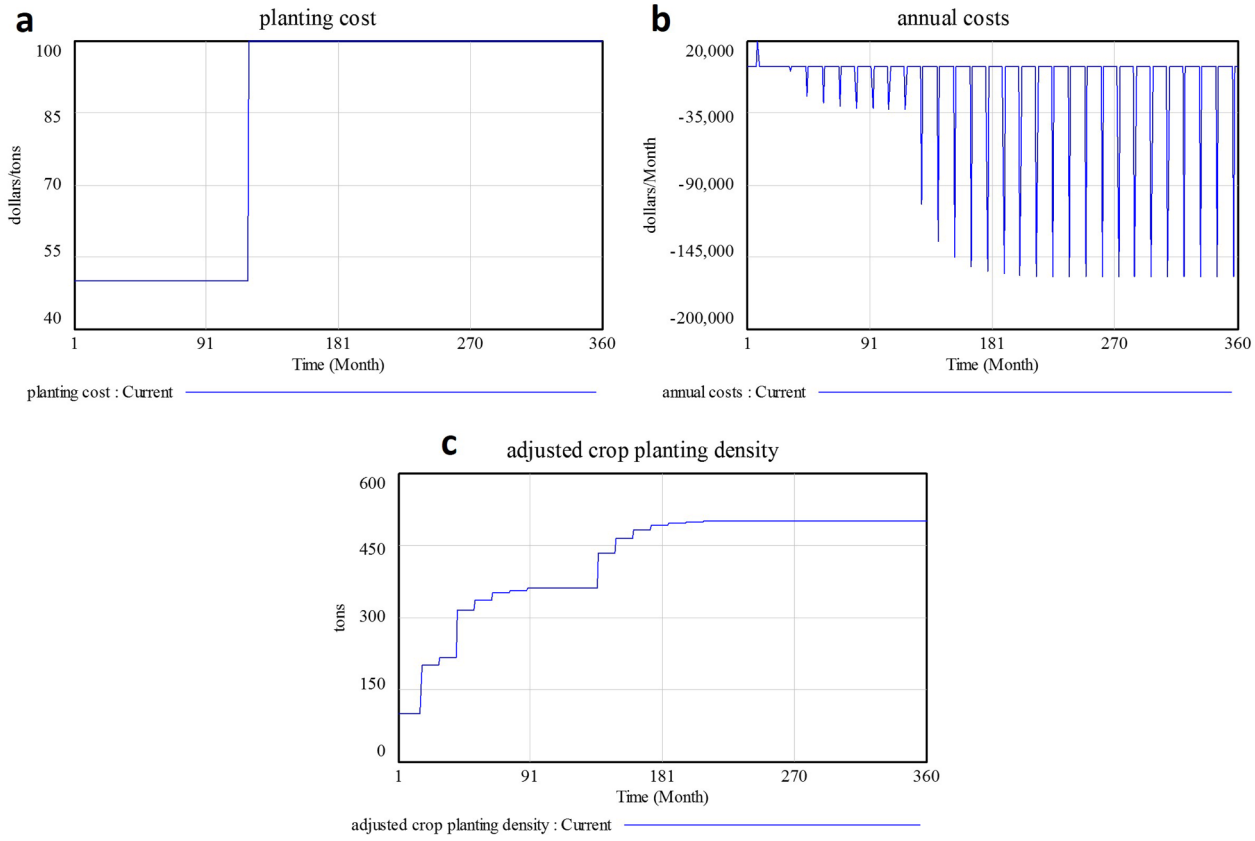

Fig. A5. Example of unexpected model response to a step function in planting costs (panel a) and the resulting decrease (into negative values) in annual costs (panel b) and the management response to increase adjusted crop planting density (panel c).

(stock = inflow – outflow), when outflows take on negative values they inadvertently *increase* the value of the stock (which in this case is the profitability of the irrigation system). Since profitability is improved rather and stressed, the adjusted crop planting density is continually adjusted up, further reinforcing crop profitability (R1 in Fig. A2). The observation that annual costs become negative is clear indication that we erroneously inserted the wrong sign, either to the cost function (flow) or cumulative profit equation (stock). Identifying and reversing the sign so that costs are positive, which have a negative effect on profitability, corrects the issue.

In Section 4.3.1., a pulse function was applied to *river flow* in order to mimic periodic drought conditions (Fig. A6 panel a) and observe the responses in *ecosystem plants* and native animals. In an earlier iteration of the test, suppose we observe that *native animals* are hardly effected by the loss of *ecosystem plants* in drought years and that they "rebound" back to near their long-run equilibrium value much too quickly (Fig. A6 panels b and c). Having previously examined the stock-flow structures around *ecosystem plants*, as well as the *natality rate* in *native animals*, we now examine the *forage availability index* (= *ecosystem plants / native animals*) influence on *mortality rate* (B1 in Fig. A2). This graphical function was parameterized using a distribution derived from data collected from the real-world system, with a mean ≈200 tons per head (corresponding to ≈14% *mortality rate*). Unfortunately, this form of graphical function breaks several best modeling practices. First, graphical functions should not start and end in the same place (in this case, at 0% *mortality rate* on the tails of each side of the curve). Because of the function is parabolic, with one side exhibiting a positive slope and another side exhibiting a negative slope, interpretation of the polarity of this variable on *native animals* is confounding, because it both accelerates and slows *mortality rate* depending on which side of the distribution the *forage availability index* value falls on. Although some circumstances may call for using a specified distribution (which many programs allow via built-in function rather than graphical/table functions), it is questionable here since the estimated *mortality rate* given the drought conditions (only 5 out of the 30 years) falls near the right and left bounds of the distribution for *mortality rate* due to sampling for *ecosystem plants* that did not accurately account for years of drought. Graphical functions can indeed be nonlinear but should have either a positive or negative slope, not both. In this case, removing the left tail (positive slope of the distribution; Fig. A6 panel e), corrects the polarity error resulting in *native animals* that express a more realistic and expected decline in behavior pattern (Fig. A6 panel f).

Not all model revisions will be in the quantitative model, but rather correct model performance that we did not properly intuit should lead us to revise our mental model. Consider a linear ramp function similar to that applied to *river flow* in Section 4.3.1., but instead applied to irrigation efficiency. In the *river flow* test, a linear increase in *river flow* led to a linear increase in crop production (Fig. 7) and profits. This was simple enough to mentally intuit, since there was no change in the rate in which water was converted into crop production only the volume of water applied as irrigation. However, applying a positively-sloped ramp change to irrigation efficiency (Fig. A7 panel a) increases the rate at which irrigation applications are converted into crop production (Fig. A7 panel b). However, because of the economic feedback between the cropping system and profitability (R1 and R2 in Fig. A2), greater crop production improves profit, signaling for greater planting densities in subsequent years, which increases demand for and applications of irrigation water that produces greater crop volumes because of the positively-sloped ramp. Intuitively, we may suspect that the linear improvement in irrigation efficiency would lead to linear increases in the cropping system. The result, however, is a nonlinear or exponential increase in water use, crop production, and profitability due to the underlying feedback processes (Fig. A2). In this particular test, the model was operating correctly and helped expose a limitation in our mental model. Updating our mental models to be more

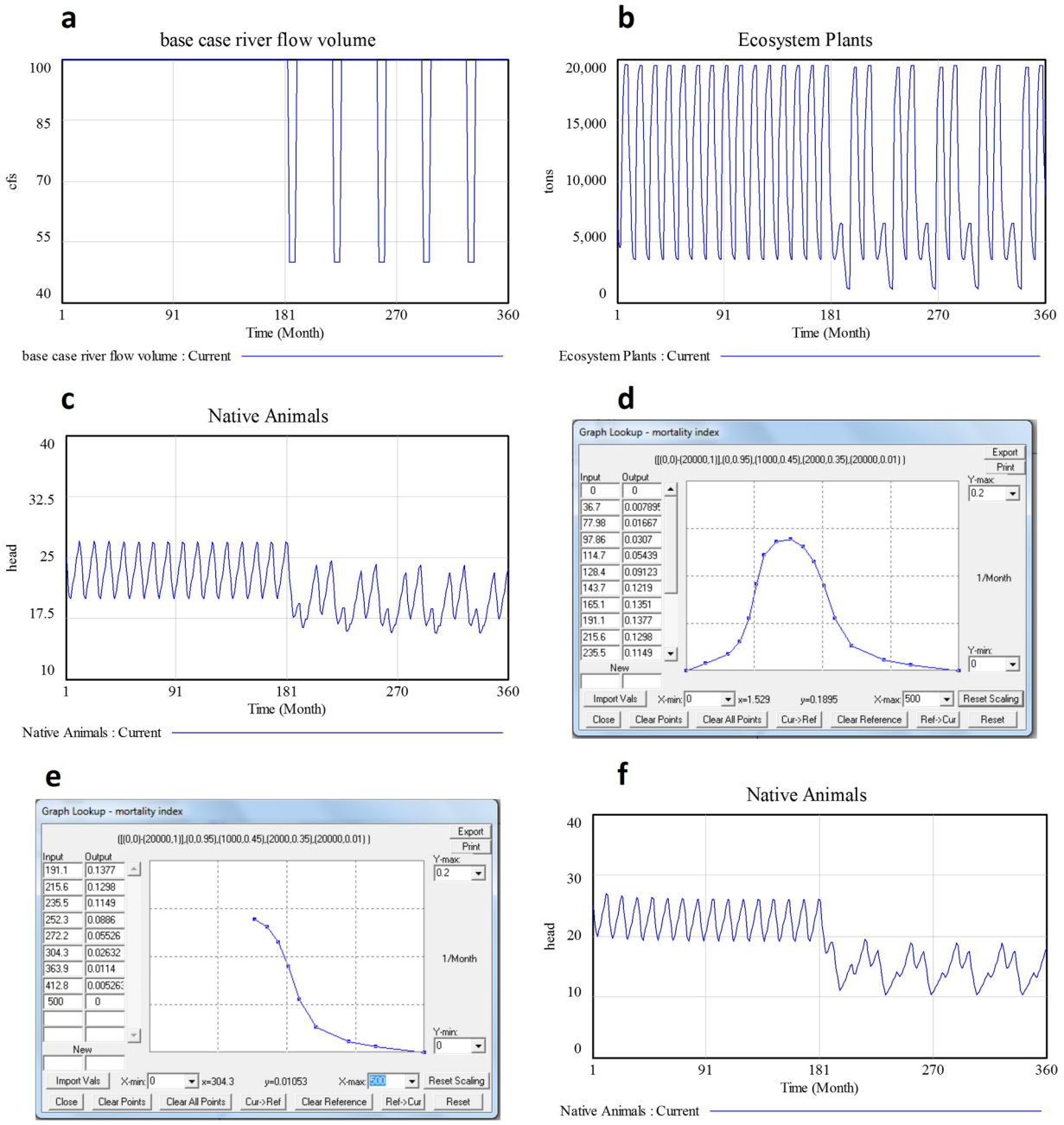

Fig. A6. Example of unexpected model response to a pulse function in river flow (panel a) and the resulting behavior in ecosystem plants (panel b) and native animals (pane c) due to an improperly reasoned graphical function for the forage availability index effect on mortality rate (panel d). One solution to the graphical function (panel e) results in the more realistic behaving decline in native animals (panel f).

representative of the dynamic and nonlinear nature of the systems we work in, although challenging, is greatly enhanced by the use of models for this reason.

3.3. Identification of behavior patterns

As described in Table 2, one of the common limitations in novice modelers is properly differentiating between alternative behavior patterns. Given the irrigation-wildlife refuge data (Fig. 3) and the atomic behavior patterns (Fig. A1), we can identify the resulting behavior patterns exhibited from the sensitivity analysis or expected patterns of behavior during intervention analysis (Fig. A8).

3.4. Comparative analysis of alternative assumptions, decision rules, or policies (counterfactuals or what-ifs?)

The final set of examples illustrate some iterative model revision steps during counterfactual trajectory analysis, boundary-adequacy testing, and intervention analysis. Counterfactual trajectories provide a means to examine model behavior under alternative basic assumptions underpinning the model about past or potential future conditions. However, the failure to adequately envision conditions that are significantly different than the original model assumptions may mislead modelers to conclude that the problem-behavior being modeled is robust enough that even under alternative conditions the same behavior patterns are observed. This could become particularly problematic when decisions will be based on insights

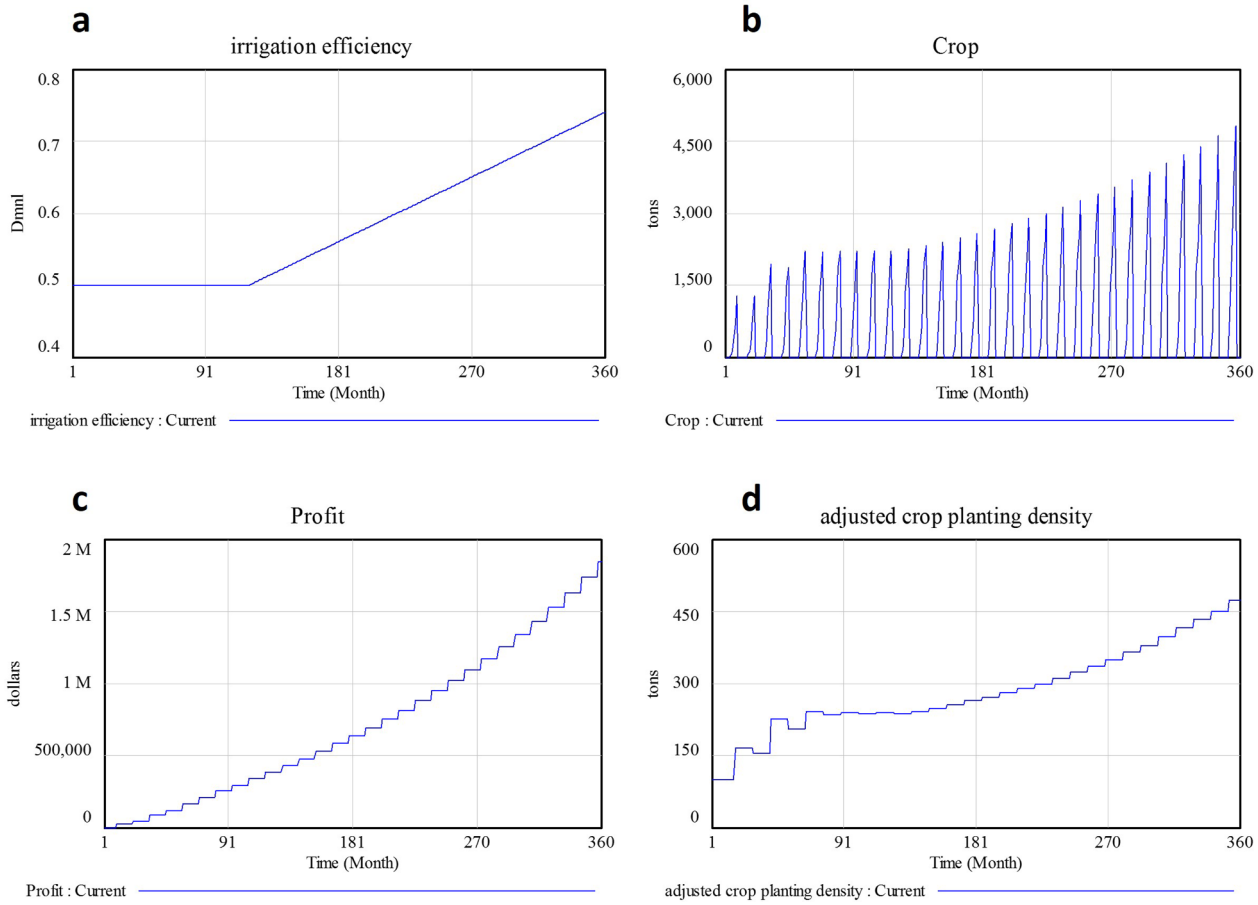

Fig. A7. Example of a correct nonlinear response in the system from a linear change in an input parameter. In this instance, irrigation efficiency (panel a) is increased linearly via a ramp function, which creates a nonlinear response in crop production (panel b), profits (panel c), and adjusted crop planting density (panel d), which reinforce each other through the economic feedback of the system (R1 in Fig. A2).

generated from counterfactual tests. For example, the counterfactual test in Section 4.4.1. represented a test for extremely variable *river flow* (distributed from 0 to 125 c.f.s.) and illustrated that both the *native animals* and the irrigation system would be highly vulnerable to long-term variability in *river flows*. However, if the counterfactual assumptions are more conservative (e.g., river from 60 to 125 c.f.s.; Fig. A9 panel a), the resulting behavior patterns in *crop production* (panel b), *ecosystem plants* (panel c), and *native animals* (panel d) would be the same as the observed conditions that motivated the study. Failure to envision a significantly new set of conditions runs the risk of basing new strategy or policy changes on flawed insights about the range of possible behavior patterns the model expresses.

Boundary-adequacy tests require expansion of the model structure to incorporate elements previously not contained in the original endogenous feedback structure of the model. This requires creativity to envision and design new model structure. Failure to do so may lead modelers to look for alternative adjustments to expand model structure. For example, a stock may be disaggregated into a series of stock-flow structures. In this case, the stock of *native animals* may be disaggregated into several age-classes that progress from younger to older individuals (Fig. A10 panel a). Unfortunately, this does not expand the model boundary, only adds specificity to model structure within the existing model boundary. Simulating the model with the disaggregated *native animal* stocks leads to an oscillatory behavior pattern in *native animals* (Fig. A10 panel b) and *ecosystem plants* (Fig. A10 panel c) which is not observed in the real-world system (Fig. 3 in Section 4.1), which may lead one to make erroneous conclusions about the adequacy of the existing model boundary and structure.

Another boundary-adequacy pitfall is likely to be insuring that, once new model structure is created, that the feedback connectivity is correctly linked with the original model structures. For example, consider the reservoir storage stock-flow structure from Section 4.4.2. Assume that the new structure is correctly formulated and upon simulation, the reservoir storage stock indeed behaves the way we would expect (Fig. A11 panel a). However, the volume of flow entering the *river refuge* (Fig. 11 panel b) remains static, and the resulting dynamics in *ecosystem plants* and *native animals* (Fig. 11 panels c and d) remain unchanged. Because the *reservoir storage* capacity is not infinite, we know that some water has to be released downstream to the refuge, so the fact that *river refuge* flow is static is a key indicator that not all of the feedback connections have been incorporated yet.

Lastly, in searching for interventions to alleviate the systemic root-cause of the problem, we can test our intuitions about strategies that would work only symptomatically in the short-term. For example, *reintroducing animals* (Fig. A12 panel a) only increases the *native animal* population for the year they are introduced, since the *ecosystem plants* required to support the population remain unchanged (panel b) as irrigation water diversion continues to support the base *crop production* (panel c). On the other hand, *supplemental feeding* for *native animals* (up to 95% of their forage demand) to alleviate pressure on *ecosystem plants* does work in the short- to medium-term (Fig. A13 panels a and b), but because revision to the water allocation mechanisms that drive both the refuge and irrigation system remain unchanged, the irrigation system (represented by *crop production* and *adjusted water irrigation level*, panels c and d) is still allowed to grow, albeit at a slower rate due to the costs of importing the *supplemental feed*. Although feeding works moderately well, the long-term result of the system is the same as if no feeding would have occurred.

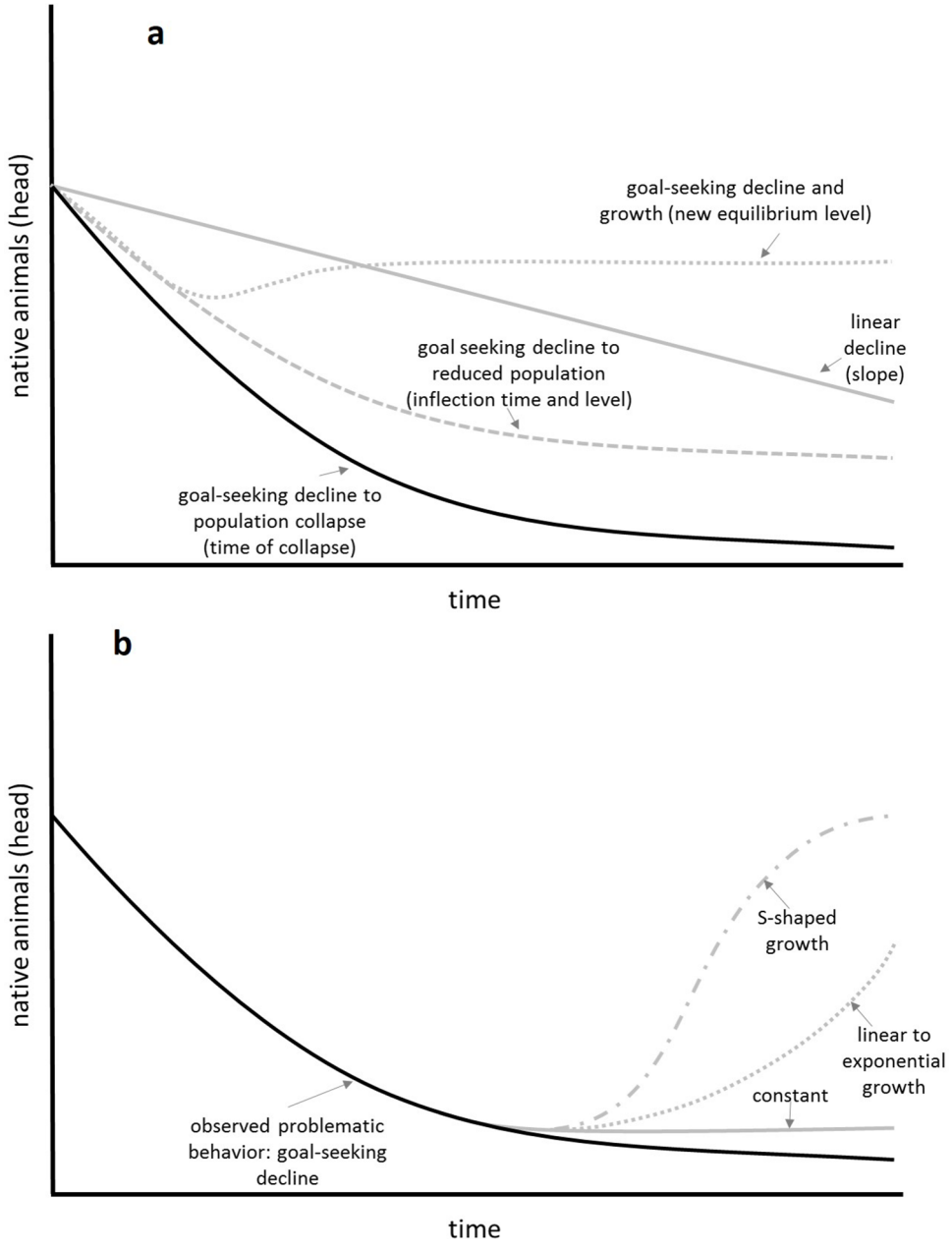

Fig. A8. Panel a illustrates the four behavior patterns expressed during multivariate sensitivity analysis. Panel b illustrates the potential behavior patterns that could be expressed during intervention thresholds analysis.

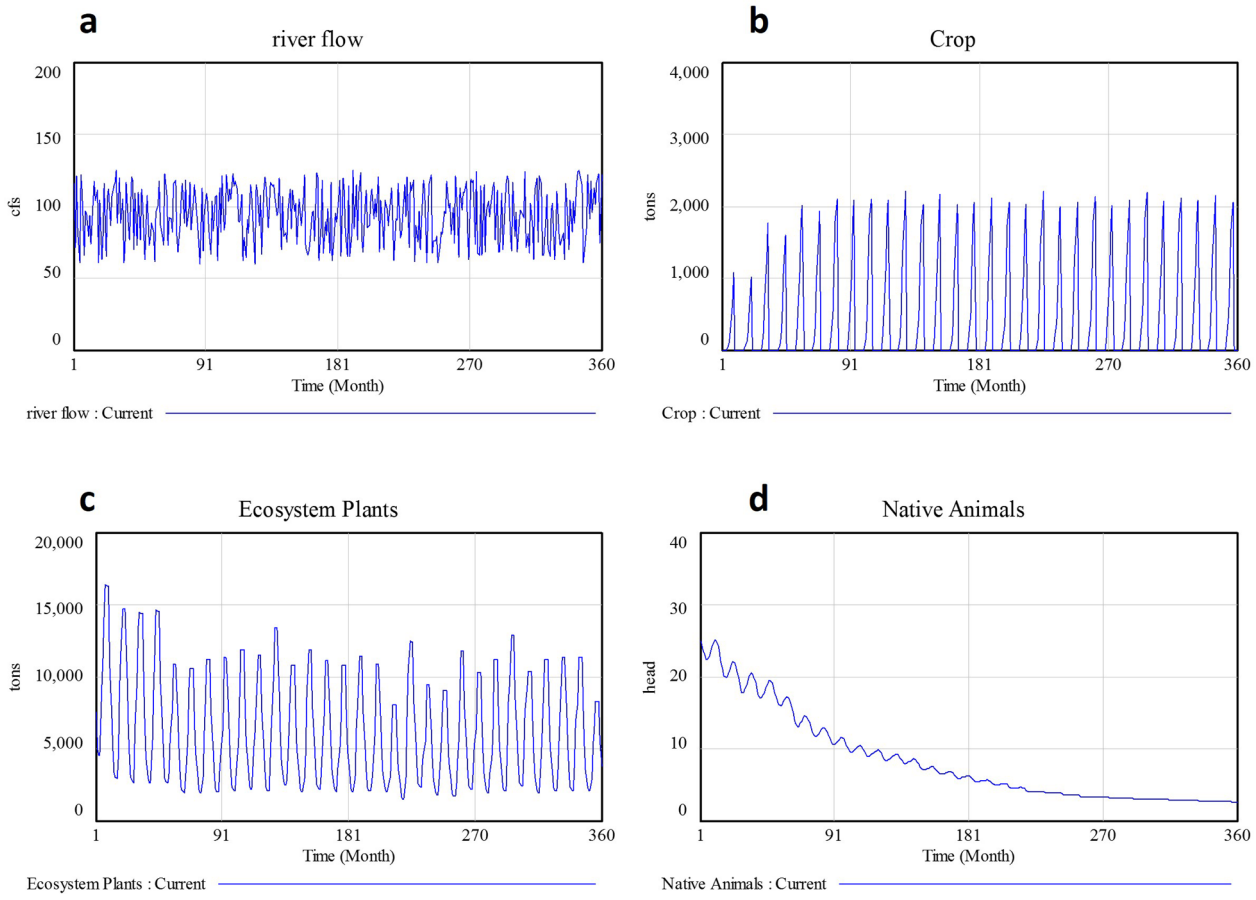

Fig. A9. Panel a illustrates a counterfactual assumption in river flow (distributed from 60 to 120 c.f.s.) and the resulting behavior pattern in crop production (panel b), ecosystem plants (panel c), and native animals (panel d).

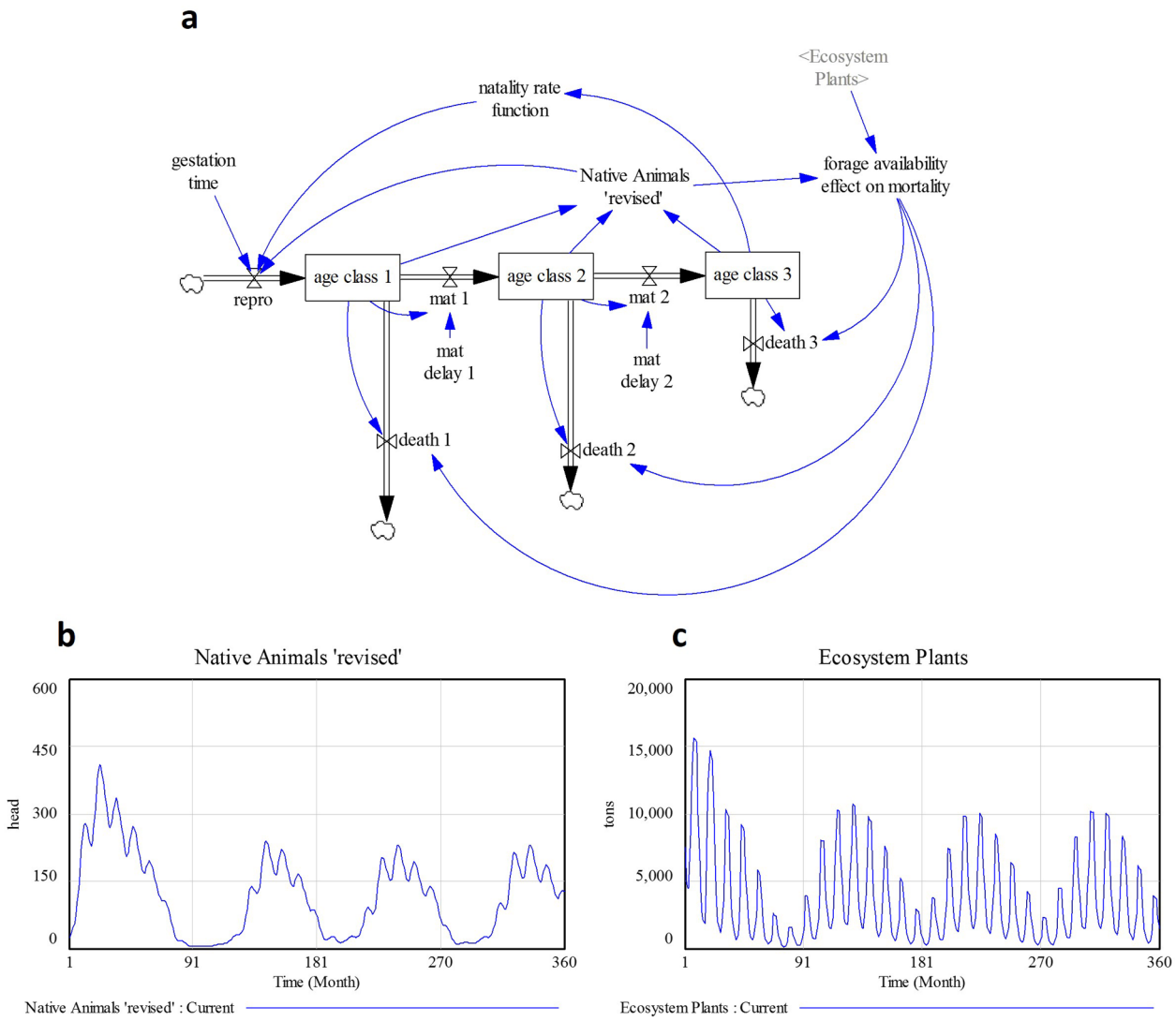

Fig. A10. Disaggregation of the stock of native animals (panel a) disguised as a model boundary adequacy test, with the resulting behavior pattern in native animals (labeled 'revised'; panel b) and ecosystem plants (panel c).

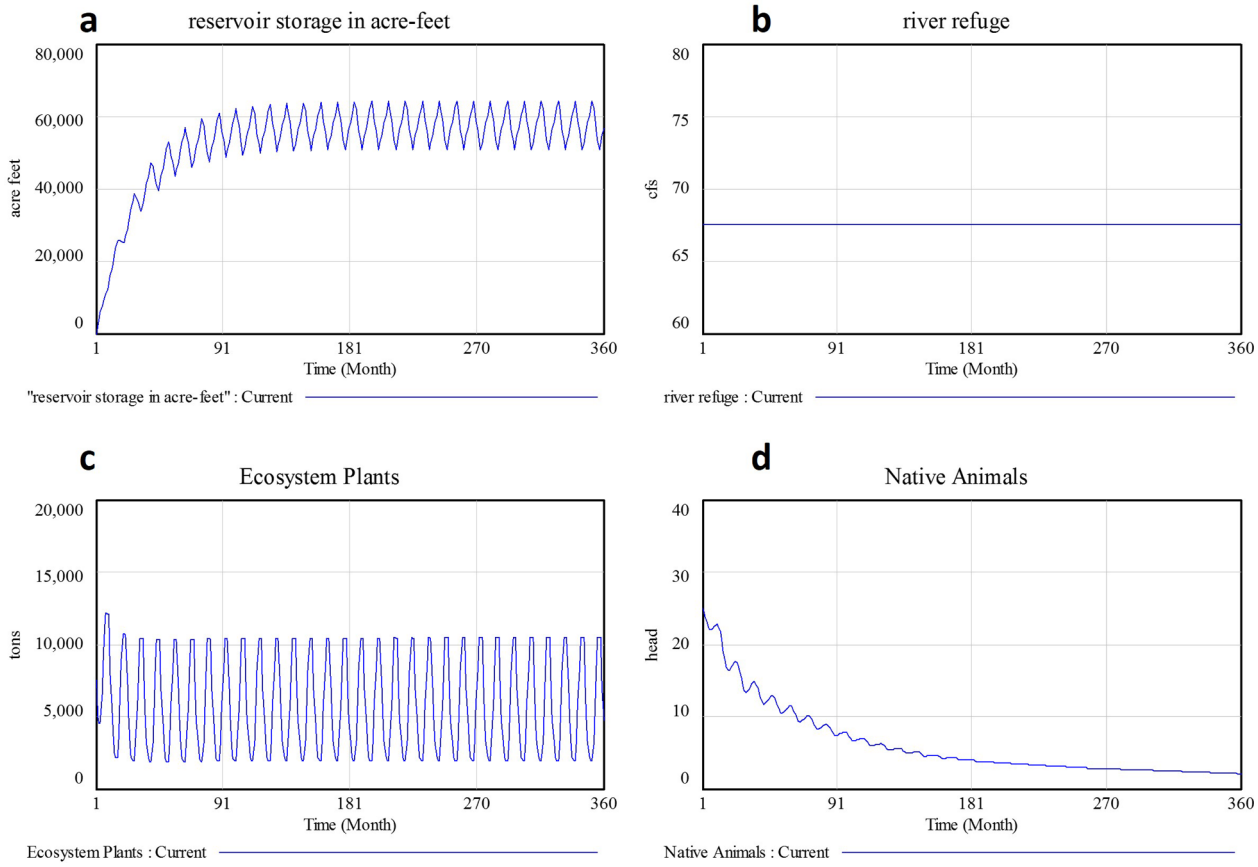

Fig. A11. Model boundary test using the reservoir storage scenario described in the paper but with an error in the water balance equations that drive water to the wildlife refuge. Reservoir storage (panel a) provides irrigation water, however the water entering the river refuge (panel b) is static, indicating that there are no return flows accounted for in this particular simulation. The reduced water to the refuge reduces ecosystem plants and therefore native animals.

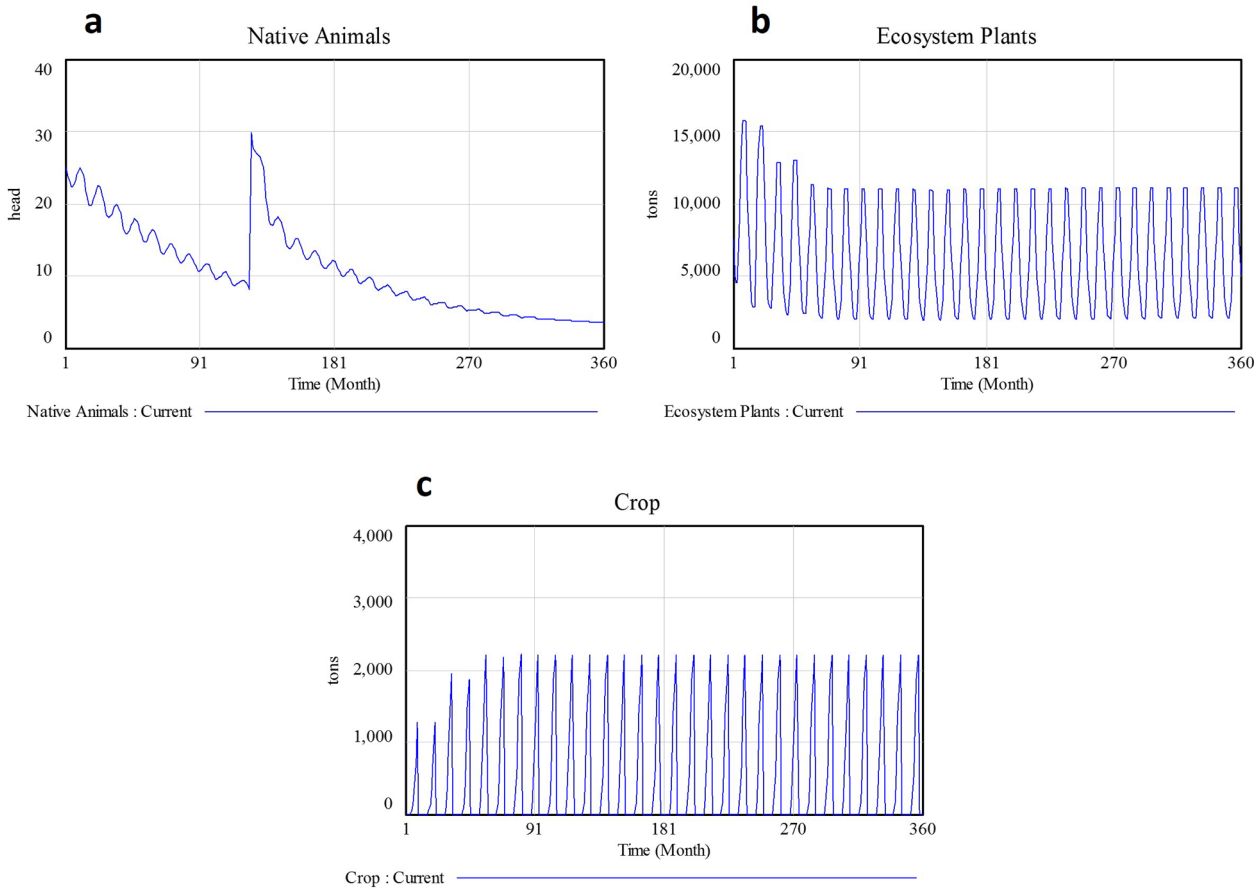

Fig. A12. Failed intervention test using native animal introduction (panel a) with no improvement to ecosystem plants (panel b) or trade-off to crop production (panel c) that drives the irrigation system behaviors.

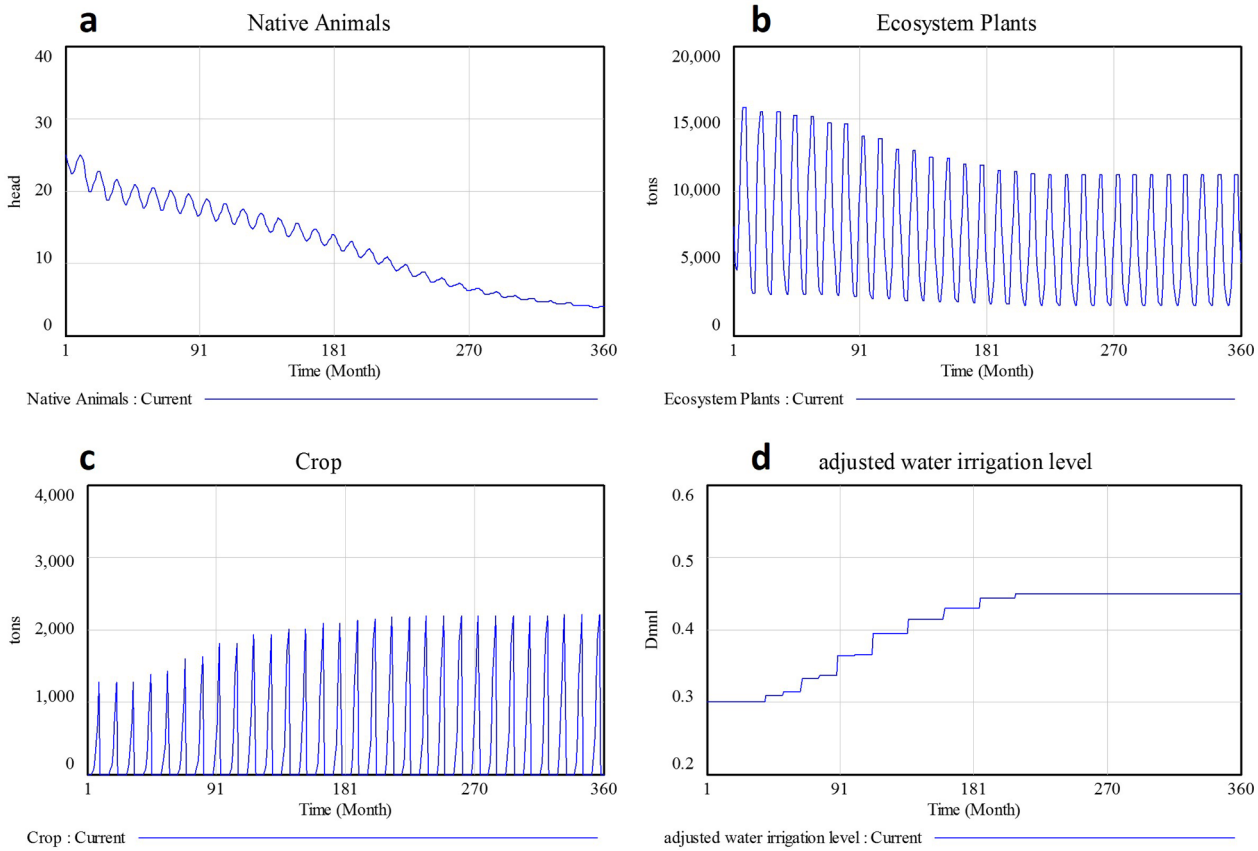

Fig. A13. Failed intervention test using feed resource supplementation (up to 95% of forage demand), illustrating the only the short-term enhancement in native animals (panel a) and ecosystem plants (panel b) and the increase in crop production (panel c) and adjusted water irrigation level (panel d) over a longer timer period relative to the original scenario.

References

Barlas, Y., 1989a. Multiple tests for validation of system dynamics types of simulation models. Eur. J. Oper. Res. 42 (1), 59–87.

Barlas, Y., 1989b. Tests of model behavior that can detect structural flaws: demonstration with simulation experiments. Computer-Based Management of Complex Systems. Spring, Berlin, Heidelberg, pp. 246–254.

Barlas, Y., 2007. System dynamics: systemic feedback modeling for policy analysis. System 1, 59.

Bennett, N.D., Croke, B.F.W., Guariso, G., Guillaume, J.H.A., Hamilton, S.H., Jakeman, A.J., Marsili-Libelli, S., Newham, L.T.H., Norton, J.P., Perrin, C., Pierce, S.A., Robson, B., Seppelt, R., Voinov, A.A., Fath, B.D., Andreassian, V., 2013. Characterising performance of environmental models. Environ. Model. Softw. 40, 1–20.

Cavazzuti, M., 2013. Optimization Methods: From Theory to Design. Springer ISBN 978-3-642-31186-4.

Cox, D.R., Reid, N., 2000. The Theory of the Design of Experiments. Chapman & Hall/CRC, Boca Raton ISBN 1-58488-195-X.

Cronin, M., Gonzalez, C., Sterman, J.D., 2009. Why don't well-educated adults understand accumulation? A challenge to researchers, educations, and citizens. Organ. Behav. Hum. Decis. Process. 108 (1), 116–130.

Dalton, G.E., 1975. In: Dalton, G.E. (Ed.), Applied Science Publishers, Essex, England.

Deaton, M.L., Winebrake, J.J., 2000. Dynamic Modeling of Environmental Systems. Springer Science + Business Media, New York, NY.

Dogan, G., 2007. Bootstrapping for confidence interval estimation and hypothesis testing for parameters of system dynamics models. Syst. Dyn. Rev. 23 (4), 415–436.

Eker, S., Slinger, J., van Daalen, E., Yücel, G., 2014. Sensitivity analysis of graphical functions. Syst. Dyn. Rev. 30, 186–205.

Ford, A., 1999. Modeling the Environment. Island Press, Washington, DC.

Ford, A., Flynn, H., 2005. Statistical screening of system dynamics models. Syst. Dyn. Rev. 21 (4), 273–303.

Forrester, J.W., 1961. Industrial Dynamics. Pegasus Communications, Waltham, MA.

Forrester, J.W., 1971/1985. The model versus the modeling process. Syst. Dyn. Rev. 1 (1), 133–134.

Forrester, J.W., Senge, P.M., 1980. Tests for building confidence in system dynamics models. TIMS Stud. Manage. Sci.14 209–228.

Grant, W.E., Pedersen, E.K., Marín, S.L., 1997. Ecology and Natural Resource Management: Systems Analysis and Simulation. John Wiley & Sons, Inc. ISBN 978-0471137863.

Gunda, T., Turner, B.L., Tidwell, V.The Influential Role of Sociocultural Feedback on Community Managed Irrigation Systems. Water Resour. Res.54(4): 2697–2714.

Hahn, G., Meeker, W., 1991. Statistical Intervals: A Guide for Practitioners. Wiley, New York.

Hearne, J., 2010. An automated method for extending sensitivity analysis to model functions. Nat. Resour. Model. 23 (2), 107–120.

Hekimoğlu, M., Barlas, Y., 2016. Sensitivity analysis for models with multiple behavior modes: a method based on behavior pattern measures. Syst. Dyn. Rev. 32 (3–4), 332–362.

Kampmann, C.E., Oliva, R., 2020. Analytical methods for structural dominance analysis in system dynamics. Syst. Dyn. 153–176.

Kelton, W.D., Barton, R.R., 2003. Experimental Design for Simulation. In: Chick, S., Sanchez, P.J., Ferrin, D., Morrice, D.J. (Eds.), Proceedings of the 2003 Winter Simulation Conference.

Kennedy, M.C., 2019. Experimental design principles to choose the number of Monte Carlo replicates for stochastic ecological models. Ecol. Modell. 394, 11–17.

Kleijnen, J.P.C., 1995. Sensitivity analysis and optimization of system dynamics models: regression analysis and statistical design of experiments. System Dynamics Review 11 (4), 275–288.

Kleijnen, J.P.C., Sanchez, S.M., Lucas, T.W., Cioppa, T.M., 2005. A User's Guide to the Brave New World of Designing Simulation Experiments. J. Comput. 17 (3), 263–289.

Langarudi, S.P., Bar-On, I., 2018. Utility perception in system dynamics models. Systems 6 (4), 37.

Leinweber, D., 1979. Models, complexity and error. A Rand Note prepared for the US Department of Energy Available at. https://www.rand.org/content/dam/rand/pubs/notes/2007/N1204.pdf.

Martinez-Moyano, I.J., Richardson, G.P., 2013. Best practices in system dynamics modeling. Syst. Dyn. Rev. 29 (2), 102–123.

Meadows, D.H., Robinson, J., 1985. The Electronic Oracle: Computer Models and Social Decisions. Wiley, Chichester.

Mohaghegh, M., Größler, A., 2020. The dynamics of operational problem-solving: a dual-process approach. Syst. Pract. Act. Res.33 (1), 27–54.

Morecroft, J.D.W., 1988. System dynamics and microworlds for policymakers. Eur. J. Oper. Res.59 (3), 9–27.

Morecroft, J.D., 1983. System dynamics: Portraying bounded rationality. Omega 11 (2), 131–142.

Naumov, S., Oliva, R., 2018. Refinements on eigenvalue elasticity analysis: interpretation of parameter elasticities. Syst. Dyn. Rev. 34 (3), 426–437.

Oliva, R., 2003. Model calibration as a testing strategy for system dynamics models. Eur.

J. Oper. Res.151 552–568.
Oliva, R., 2015. Linking structure to behavior using eigenvalue elasticity analysis. Anal. Method. Dyn. Model. 207–239.
Oliva, R., 2016. Structural dominance analysis of large and stochastic models. Syst. Dyn. Rev. 32 (1), 26–51.
Oliva, R., 2019. Intervention as a Research Strategy. J. Oper. Manag.65 710–724.
Pearce, S.C., 1983. The Agricultural Field Experiment, a Statistical Examination of Theory and Practice. John Wiley & Sons, Chichester.
Peck, S.L., 2004. Simulation as experiment: a philosophical reassessment for biological modeling. TRENDS Ecol. Evol. 19 (10), 530–534.
Peterson, D.W., Eberlein, R.L., 1994. Reality Check: a bridge between systems thinking and system dynamics. Syst. Dyn. Rev. 10 (2–3), 159–174.
Rahmandad, H., Sterman, J., 2008. Heterogeneity and Network Structure in the Dynamics of Diffusion: comparing Agent-Based and Differential Equation Models. Manage. Sci. 54 (5), 998–1014.
Repenning, N.P., 2001. Understanding firefighting in new product development. J. Prod. Innov. Manage.18 (5), 285–300.
Richardson, G., Pugh, A.L., 1981. Introduction to System Dynamics Modeling with DYNAMO. J. Oper. Res. Soc. 48, 1146.
Sanchez, S.M., 2005. Work Smarter, Not Harder: guidelines for Designing Simulation Experiments. In: Kuhl, M.E., Steiger, N.M., Armstrong, F.B., Jones, J.A. (Eds.), Proceedings of the 2005 Winter Simulation Conference.
Saltelli, A., Chan, K., Scott, E.M., 2000. Sensitivity Analysis. Wiley, Chichester.
Srinivasan, V., 2015. Reimagining the past – use of counterfactual trajectories in socio-hydrological modeling: the case of Chennai, India. Hydrol. Earth Syst. Sci.19 785–801.
Sterman, J.D., 1994. Learning in and about complex systems. Syst. Dyn. Rev. 10 (2–3), 291–330.
Sterman, J.D., 2000. Business Dynamics: Systems Thinking and Modeling for a Complex World. Irwin/McGraw Hill, Boston ISBN 0-07-231135-5.
Sterman, J.D., 2002. All models are wrong: reflections on becoming a system scientist. Syst. Dyn. Rev. 18 (4), 501–531.
Sterman, J.D., 2009. Does formal system dynamics training improve people's understanding of accumulation? Syst. Dyn. Rev. 26 (4), 316–334.
Tank-Nielsen, C., 1980. Sensitivity analysis in system dynamics. In: Randers, J. (Ed.), Elements of the System Dynamics Method Productivity Press: Cambridge, MA.
Tedeschi, L.O., 2006. Assessment of the adequacy of mathematical models. Agric. Syst. 89, 225–247.
Turner, B.L., Menendez, H.M., Gates, R., Tedeschi, L.O., Atzori, A.A., 2016. System Dynamics Modeling for Agricultural and Natural Resource Management Issues: review of Some Past Cases and Forecasting Future Roles. Resources 5 (4), 40.
van Belle, G., 2002. Statistical Rules of Thumb. Wiley, New York.
Walrave, B., Van Oorschot, K.E., Romme, A.G.L, 2011. Getting trapped in the suppression of exploration: a simulation model. J. Manage. Stud. 48 (8), 1727–1751.
Walrave, B., 2016. Determining intervention thresholds that change output behavior patterns. Syst. Dyn. Rev. 32 (3–4), 261–278.
Yücel, G., Barlas, Y., 2015. Pattern recognition for model testing, calibration, and behavior analysis. In: Rahmadad, H, Oliva, O, Osgood, ND (Eds.), Analytical Methods for Dynamic Modelers. MIT Press, Cambridge, MA, pp. 173–206.