Appendix A: A historical perspective of system dynamics models validation

From this seminal book Forrester *(1961)* three main criteria can be extracted regarding the validity of a system dynamics model: system boundaries, interacting variables, and values of parameters. System boundaries are the most important criteria. Choosing a small boundary creates a system without the endogeneity necessary to understand loops. An oversized system might be distracting and lead to confusion and abandonment of the model. Interacting variables refer to the question if the model embraces the relevant variables and if they are adequately connected. Here, the challenge is to understand if the list of variables used in the model is relevant to the system simulation, but mostly to understand how these variables interact with each other. In both cases, experience is the best guide *(Forrester, 1961)*, which reinforces the idea of modeling being a process of learning and experimentation. Values of parameters are the least important aspect contributing to the validity of a model. System dynamics models are usually lowly sensitive to values variation, and these constant values are only statistically tested after broader and deeper assumptions were already used in the model, such as: objectives were decided, boundaries were determined, relevant variables were chosen, a hypothesis of how each variable interact was created, and an arbitrary level of confidence was established for the statistical test.

The work of Forrester and Senge *(1980)* is still a main reference regarding a structure for testing the validity of models in the system dynamics field. It is based on three categories of tests:

1. test of model structures - before testing the behavior and outputs of the model, this step of testing will focus on the structure of the model and the system it is simulating;
2. test of model behavior - these tests are focused on the outputs of the model, usually numeric, that can be forecasted or backcasted for a specific attribute of the system. Tests of model behavior should be applied when the model at stake can produce any type of forecasting or backcasting that allows testing; and
3. tests of Policy Implications - These kinds of tests are a comparison of the system behavior after the application of policies tested in the model, with the outputs forecasted in the simulation.

Meadows (1980) brings a deeper understanding to the discussion, highlighting the relevance of *a priori* assumptions each modeler translates to the model, conscient or not.

The author claims an analyst starts to trust a model when it meets the following conditions:

1. Every element and relationship in the model have identifiable real-world meaning and is consistent with whatever measurements or observations are available.
2. When the model is used to simulate historical periods, every variable exhibits the qualitative, and roughly quantitative, behavior that was observed in the real system.
3. When the model is simulated under extreme conditions, the model system´s operation is reasonable.

The validation tests according to Barlas *(1989, 1996)*, can be categorized into two main domains: Structural and Behavioral tests. Despite being highly cited in the field, he is mostly concerned with the validation of numerical models and therefore escapes the limits of the present piece. Nonetheless, in the middle of the 1990s, Barlas understood that there was a likely minor complexity type of models, which he refers to as models-for-learning *(Morecroft & Sterman, 1994)*, in contrast to models to "improve performance". These learning models fit into his understanding of system dynamics in the same way as theory-testing models do, meaning a group of simulations with less rigor, broader participation, focus on learning and experimentation, and finally, with a lower necessity of thoroughly testing: "the models built for learning may not necessitate such behavior accuracy testing as the traditional applications do" *(Barlas, 1996)*.

The contribution of Balci *(1994)* comes manyfold. The author conceptualized the whole modeling process in a timely life-cycle perspective. In addition, defined the main types of errors the modeling process can produce, such as: type I – rejecting the model due to lack of quality when it has enough quality to be accepted; type II – accepting the model credibility when in fact it is not sufficient credible; and type III – solving the wrong problem (which corroborates the relevance of scope and boundaries definition beforehand or problem formulation in Balci´s terms). Balci´s test set is divided into five categories (informal, static, dynamic, symbolic, constraint, and formal), disposed of in a crescent level of formality, embracing 45 individual tests, from which we selected the contributions that are timely, and which would be relevant to the present study (Appendix A). Several of these tests are overlapping or with minor variations, with each other and with those provided previously *(e.g., Forrester & Senge, 1980)*. Some of the tests, for example, the "review test" is so broadly defined that embraces a whole set of modeling tests. Nevertheless, one of its recommendations regards the timely idea of documentation,

and therefore a new line in the test matrix (Appendix A) was created with this specific test.

Another remarkable contribution to system dynamics modeling and validation is the Folding Star Framework *(Lane, 1995)*. In his schematic, the author provides guidance on the system dynamics modeling process based on the following steps: first, an AoS (Appreciation of the Situation) which can be understood as part of the question formulation and some considerations about the status of the system. This part is based on the works in soft systems *(e.g., Checkland, 1989; Checkland & Poulter, 2020)* with the division of the real system into natural, designed physical, and designed abstract systems (from which cultural attributes emerge and can be considered in the analysis). Second, from the AoS, a Communicated Conceptual Model (CCM) is produced, representing a qualitative representation of the situation, with the clear objectives of sharing the views and problems about the situation with other participants, comparing these multiple views, to finally building an understanding about the system. This step can be used to underpin a mathematical model, the Formal Model (FM). Although this step is recommended by the author, it is not mandatory and the CCM is understood as having enough legitimacy and utility to underpin the final objective of the exercise; third, the creation of the Policy Insights or Recommendations (PIoR), which brings the results produced during the exercise in terms of recommendations to change the system, closing the qualitative loop of the folding star.

This qualitative loop of the star shows three levels for validation: conceptual, inferential, and operational validity. Conceptual validity refers to the coherence of the CCM regarding the AoS, where the ideas of the community about the system must be seen in the model, including cultural aspects (an appreciation of values and ideas the group believes are worth pursuing). Inferential validity measures the extent to which the PIoR can be deducted from the CCM. This is the most fragile part of the scheme and has been a target for criticism (see the limits and caveats topic) due to the big leap required from a CCM to PIoR (in the absence of an FM). Finally, the operational validity is related to the influence the model has in the AoS, forming the feedback of the process into the understanding of the situation. This operational step is more developed for the FM, but regarding the CCM the main ideas concern the realism of the model, the analytical quality of the PIoR, and the satisfaction felt with the process. In general intermediate-quality states are considered reasonable targets due to the lower analytical potential of qualitative models *(Lane, 1995)*.

Some remarkable aspects of the folding star to the present piece are, first, to bring to the discussion a group of variables related to cultural assumptions in the model (naturally qualitative). The cultural assumptions tests concern the social elements of an system dynamics activity. These elements of investigation try to embrace the "different perceptions of a problem that might exist in and to address the social realities of the group" *(Lane, 1995)*. To the author, exposing the differences of opinion regarding the problem is crucial before converging into the problem statement. As the AoS and CCM are plurally discussed, a deeper appreciation of values and other cultural aspects embedded in the model is expected, possibly resulting in meaningful recommendations to PIoR. Second, it brings a set of tests regarding the usefulness of the modeling process, including but not limited to, the costs and time involved, the social/political capacity to implement the recommendations, the affinity between the participants with the modeler and with the modeling process, amongst others (Appendix A). This meta-understanding of the modeling process might bring relevant information regarding the satisfaction of those involved in the modeling exercise and crucial aspects related to the social practice enabling or obliterating the implementation of the recommendations discovered by the simulation process.

These ideas of broadening the participation in problem formulation had their roots probably with the foundation of operational research with Churchman et al. *(1957) (Reisman & Oral, 2005)*, but during the 1970-1980s, they were markedly reinforced and formed the emerging field of the soft systems methodologies (SSM). This school of thought is concerned with the plurality of social participation in the modeling, as stated: "What is the system? What are its objectives? ignore the fact that there will be a multiplicity of views on both, with alternative interpretations fighting it out on the basis not only of logic but also of power, politics, and personality." *(Checkland, 1989)*. The main point defended by the author is that Systems Analysis, Systems Engineering, and Operational Research, despite small variations, deal with the same thing, namely well-defined problems. To these problems, the elegant solution (i.e., optimal or efficient), suits the goals of the modeling process. On the other hand, SSM goes for a messy, ill-defined, frequently contested terrain, where the elegant solution rarely will be the answer since conflicting worldviews act upon the definition of the problem and the solution. This broader understanding is congruent with more modern views of system dynamics, considering there are no value-free theories and no value-free models *(Sterman, 2000)*.

Another important trait in SSM that makes this perspective suitable for social-ecological systems modeling is that the problem is never taken as permanently solved. The solutions obtained, by a process of accommodation of those conflicting worldviews, are always provisional since the group of assumptions that were considered in that negotiation can change, namely, the state of the system, the balance of power, the emergence of new problems, or even the worldviews *(Checkland & Poulter, 2020)*. The main objectives of SSM then are to organize the process of discussion towards the solution of a problem (purposeful action), in a constantly learning perspective, that produces solutions both desirable (in terms of the options given/structured by the decision-making process) and feasible (meaning tolerable, considering the conflicting worldviews). From this perspective of SSM, more recent branches of systems thinking emerged, such as: holistic flexibility *(Chowdhury et al., 2023)*.

The quality assurance process of a SSM, is very basic, embracing the ideas of: efficacy – whether the transformation in the system is producing its intended outcome; efficiency – whether the transformation is being achieved by using the minimum amount of resources; and effectiveness – if the transformation is helping to achieve a long-term or higher-level aim *(Checkland, 1989)*. Other elements can be added to that, such as elegance, understood as an aesthetic criterion, differing from that optimal criteria as in system dynamics; or ethicality, which questions the ethical foundations of the transformation proposed *(Checkland & Poulter, 2020)*.

In a broad sense, the author recommends users check the model for coherence, and uses the term defensible, instead of correct, to name a model which passed this coherence test, to which extent every connection in the model is meaningful in terms of understanding the Root Definitions (boundaries) and the CATWOE (a mnemonic for the process of modeling, meaning C: costumers, A: actors, T: transformation, W: worldviews, O: owner, E: Environment) *(Checkland & Poulter, 2020; Haynes, 1995)*. These three elements of quality assurance (efficacy, efficiency, and effectiveness) were understood as being congruent to the system improvement, time-and-cost of investigation, and family member tests respectively (Appendix A), to which the worldviews/values and boundary-related tests were added considering the central ideas in SSM.

To the limits of our knowledge, the issue of validation/verification of system dynamics models got cold during the last 20 years, with much fewer publications. Exceptions made by some articles *(e.g., Lemke & Łatuszyńska, 2013; Qudrat-Ullah, 2012; Qudrat-Ullah & Seong, 2010)* that reproduced Forrester´s 1980´s ideas with the

improvement of the mathematical approach for quantitative tests, but with minor theoretical increment. The work of Andersen and collaborators *(Andersen et al., 2012)* called attention due to the focus on validating qualitative models. In their contribution, disconfirmation (i.e., invalidation) of a causal construct can be made by asking a third party, not involved in the modeling section, to create a judgment about the model, where the level of agreement with the simulation corresponds to the level of validity of the model. We argue this approach can only make sense if the worldviews and values of both people interviewed during the model building and disconfirmation stage are congruent. If we assume building causal models are similar to a theory creation *(Forrester, 1961; Sterman, 2000)*, and worldviews are determinant of the way people frame their understanding of the world and its problems *(Ney, 2012; Thompson et al., 1990; Verweij et al., 2006)*, it would be expected that people with conflicting worldviews would disconfirm previous models not based on their validity to simulate a problem, but due to the differences in how they frame problems. That was already recognized by part of the system dynamics community *(Checkland, 1989; Checkland & Poulter, 2020)*.

The most recent review *(Schwaninger & Groesser, 2020)* brings some important elements of validation, including a useful loop-dominance idea (Appendix A), but also ignores SSM and the debate about embedding multiple rationalities in modeling. It is also more dedicated to the validation of quantitative simulations, following the traditional approach *(e.g., Forrester & Senge, 1980)*.

The foundations of presumed utility

Causality models represent a form of a theory of how a system works. It describes the connections of the elements of the system in such a way that one can create an understanding of the system by understanding the causalities described in the model. On the other hand, statistical models are based on ideas of correlation between variables in the system that can be used to forecast or predict the behavior of a system, preferably inside the same parameter range to which the correlation was observed, but cannot produce the same explication feature offered by a causal model *(Barlas & Carpenter, 1990)*. This movement can be understood as an analogy of the evolutionary perspective of the 1950-1970s scientific transition in which “the positivists' earlier preoccupation with "prediction," which they had regarded as the key evidence of scientific `knowledge,´ was being supplemented by a concern with `explanation,´ regarded as the core of scientific

`understanding´ *(Radzicki, 1990; Toulmin, 1977)*. This methodological/philosophical dichotomy has been tested recently and showed that the predictive capacity of both causal (deductive) and statistical (inductive) models are equivalent with the advantage of causal models providing additional explicative power to the results *(Overmars et al., 2007)*.

A brief review of the underlying philosophies concerning scientific theory development will point out that the dichotomy in these branches is deeper than purely methodological. Authors claim there are philosophical schools that might justify the division shown above and additionally underpin the assumptions of validation of theory/models, namely: the empiricist/reductionist school and the relativist/holistic school *(Barlas & Carpenter, 1990)*. This division into two main branches in the philosophy of science was corroborated in Economics studies. although with distinct names, where the empiricist/reductionist was called neoclassical (logical empiricism) and relativist/holistic named institutionalists (pragmatic instrumentalism) *(Radzicki, 1990)*.

The empiricist/reductionist school is a perspective rooted in Kant's epistemology and based on ideas of knowledge being entirely objective, ahistorical, asocial, and acultural, to which an absolute truth can be reached independently of human values and belief *(Barlas & Carpenter, 1990)*. This school, posteriorly discussed by Russel, the early Wittgenstein, and the contributions of the Vienna Circle, evolved to the logical empiricism of early 20th century, which focused on the reduction of scientific statements to the criteria of being validated by the direct observational statements, with great rigor in the meaning of each statement, and avoiding ambiguities, vagueness, and inconsistencies. Popper collaborated with this school in his early days, by advocating the criteria of falsifiability (instead of verifiability) to which a theory gains trust as much as failed attempts to prove it wrong accumulate. From this falsifiability idea comes the statement that every piece of knowledge is fallible because its status of valid is always provisional due to the possibility of being falsified by new evidence. One of the main critiques here comes from the Kuhnian perspective of scientific knowledge being biased by the Normal ruling paradigm, and consequently heavily historically and socially influenced *(Kuhn, 1962)*.

The relativist/holistic school is a perspective rooted in Hegel´s epistemology, where scientific ideas are byproducts of an Age, influenced directly or indirectly by the social foundations in which it was created. Knowledge is, thus, socially, historically, and culturally dependent *(Latour, 2013)*, therefore there cannot be a neutral foundation, and a pure objective verification is not possible *(Barlas & Carpenter, 1990)*. From this school,

the idea of pure knowledge, independent of social and historical processes are abandoned towards interdisciplinary more flexible ideas, where the absolute truth out of formal rigor, opens space for more functional perspectives: "The academic `soundness´ required rigor of a kind that, in these gray interdisciplinary areas was simply not there to be had" *(Toulmin, 1977)*.

This relativist/holist perspective is convergent with other theories in complexity science that represents the vanguard of interdisciplinary knowledge when tackling problems of the 21st century. These recent approaches understand the part of complex problems regarding society as wicked or messy (instead of tame) *(DeFries & Nagendra, 2017; Ney, 2012; Rittel & Webber, 1973; Verweij et al., 2006)*, to which policy problems are socially defined, and therefore dependent on the plurality of views society produces (see plural views below). Moreover, when considering a plural society, public goods are in dispute, meaning public policies cannot be correct or false, they are always dependent on each social group these policies are representing. Public policies cannot propose an "optimal solution" since what is optimal for one group, might be the obliteration of others. This perspective sees the boundaries of the pressing problem as becoming less clear-cut as the connectivity of global society increases, and thus, far more dependent on framing, debate, and controversies, byproducts of a plural society. In short:

> "…in a pluralistic society, there is nothing like the undisputable public good; there is no objective definition of equity; policies that respond to social problems cannot be meaningfully correct or false; and it makes no sense to talk about "optimal solutions" to social problems unless severe qualifications are imposed first." *(Rittel & Webber, 1973)*

A plural view of society, as described by the theory of plural rationalities *(Douglas & Wildavsky, 1983)*, understands the scientific work as biased by worldviews, despite the claims of neutrality of the scientific community. The basic assumption is that social relations provide the individual with normative and cognitive tools to understand the world (shared values and beliefs). Here, politics, decision-making, technology, and social choice are understood as dependent on cultural backgrounds, described by shared values and beliefs (or worldviews), to which a typology framework *(Schwarz & Thompson, 1990; Thompson, 1997)* is proposed. With this framework, conflicting perspectives about the pressing problems can be understood and managed by a conflict-reducing heuristic *(Ney, 2012; Oliveira, 2022; Scolobig et al., 2016)*.

These interdisciplinary approaches are complemented by the perspective of Post-Normal science *(Funtowicz & Ravetz, 1997)* since it provides an understanding of science with greater openness to democratic participation. In post-normal science, "facts are uncertain, values in dispute, stakes high and decisions urgent" *(Funtowicz & Ravetz, 1993)*. This perspective enhances the idea of stakeholder consultation to broader and deeper participation, expanding the conception of stakeholders, usually restricted to the scientific community plus relevant decision-makers, to the broad community (expanded peer community) based on the justification of shared risks of the globalized civilization.

The relevance of this discussion to the validation of system dynamics models is massive and two-fold. First, it is a justification after many criticisms from the mathematical (empiricist/reductionist) "pure view" of scientific models about the criteria used in the field *(Forrester et al., 1974; Forrester & Senge, 1980; Nordhaus, 1973; Radzicki, 1990; Zellner, 1980)* far from being exhaustive. But second, a pure statistical validation process, if restricted to a mathematical formal test, is far from delivering the desired comprehensiveness of a quality enhancement process toward a useful (in terms of the presumed utility) model. Therefore, if one considers an empiricist/reductionist perspective, validation should be done strictly via a formal mathematical process, to which the result would be Boolean (true or false). Validity then, becomes a matter of formal accuracy, rather than practical use *(Barlas & Carpenter, 1990)*. On the other hand, when a relativist/holist perspective is adopted, which is the most appropriate to the present case, the validation of the model becomes something dialogical, iterative, and a process towards learning and participation. In this perspective, models are not necessarily true or false, but open to the new axis of usefulness, under the limitations of a partial, provisional, and socially accepted validity. Here, no model can claim absolute objectivity since every model carries in it the modeler´s worldview *(Barlas & Carpenter, 1990)*.