Appendix F: Presumed Utility Protocol dimensions and criteria

Dimension 1: Guidelines and process

This dimension embraces indicators 1 to 7 (Table 1) and has 85 comments (Appendix C). The average results varied from 3.3 (boundary adequacy) to 4.1 (Meaningfulness of the process). This dimension focuses on the initial steps of the modeling process, and it had many more positive evaluations (50%) when compared to negative ones (6.3%). This is a positive result because the time dedicated to model preparation and group model construction was short (as commented by several participants).

Criterion 1 (Purpose) (average 3.9, mode 4) had twelve comments, with the majority pointing that a clear understanding of the purpose was made prior to the modeling exercise. Purpose is one of the most central criteria of a modeling exercise (Schwaninger & Groesser, 2020; Sterman, 2000), and when it is not satisfactorily evaluated, indicates that the group of modelers are not sure about their tasks, which might require a review of the model or even the necessity of rebuilding if the users also report a lack of meaning (see criterion 7). Recommendation: more time discussing the purpose of the model.

Criterion 2 (Usefulness) (average 3.6, mode 4) had thirteen comments, that revealed a concern of the modelers on how to communicate the content of the model to a lay and broader audience. Usefulness can be improved by considering the level of scientific training of the audience (Sterman, 2000). Recommendation: a previous introduction of the audience to the problems, definitions, and delimitations that the model is dealing with and some instructions on how to navigate a CLD.

Criterion 3 (Presentation) (average 3.5, mode 3) had thirteen comments that showed a spectrum of different understandings, going from the idea of the model being easily understandable and intuitive to the necessity of great simplification of the ideas and reducing the number of loops. We echo Balci (1994) that the presentation should be made concerning the intended use of the model, and therefore the recommendation is to modulate the CLDs to simplify the presentation to individual loops (instead of the whole model) and associate the presentation with other techniques that help communicate with the audience (such as maps, graphs, pictures, or others).

Criterion 4 (Perspectives in Boundary-adequacy) (average 3.3, mode 3) the twelve comments on this criterion showed the adequacy of the modeling process in

supporting the debate of different perspectives, but modelers also identified the lack of variety within their group (their own bias). As extensively described and explored in the soft systems methodology models (Checkland, 1989; Checkland & Poulter, 2020), these different perspectives and goals can be explored as they provide complementary views of the system. The recommendation from the modelers is to include more people with different perspectives in the modeling exercise (other sciences and stakeholders).

Criterion 5 (Norms/values in boundary adequacy) (average 3.8, mode 4) presented nine comments revealing a great satisfaction with the capacity of the modeling exercise to promote "organized and polite" discussions along the modeling process, but the bias of the modeling group (and the necessity of interdisciplinary participation) was a constant. The modeling process was expecting some conflicting perspectives regarding the problems in each region due to different worldviews (Douglas, 2013; Oliveira et al., 2024; Thompson & Verweij, 2004), but apparently, the differences among the modelers were not sufficient to require the construction of multiple models to capture these multiple views (e.g., Checkland & Poulter, 2020). The recommendation is to include people with different perspectives in the exercise.

Criterion 6 (Trustworthiness or Guru status of the system dynamicist) (average 4, mode 0) The eleven comments here show that some positive relations with stakeholders were identified, but as the model was not presented to each area stakeholder, most of the respondents pointed that this item does not apply. As Sterman (2000) explains, this positive relationship can promote a better understanding of the modeling process, finally helping with a mutual understanding between the modeler team and the stakeholders. Recommendation: increase the connections with their stakeholders to promote better connections (by discussing these models for instance).

Criterion 7 (Meaningfulness of the process) (average 4.1, mode 0) The thirteen comments come twofold. Some modelers understood that criteria refer to "third party stakeholders", and then evaluated it as not applying, as they did not discuss the model with stakeholders yet. For the group of modelers that understood they were the stakeholders this criterion refers to, the evaluation was very positive as they saw the meaning of the process and the model (Appendix C, Appendix D question A). We echo Lane (1995) that meaning is promoted when assumptions are made explicitly in the model, and in the present case, the modelers experienced that through the workshops and materialized these discussions with the model. Recommendation: show the model to stakeholders and other relevant actors as they were absent so far in the process.

Dimension 2: Specific model tests

This dimension embraces indicators 8 to 12 (Table 1) and has 59 comments. This dimension had more positive results (55%) than the sum of all other categories of answers (Figure 5). The average results varied from 3 (extreme conditions) to 4.1 (boundary adequacy). The first was known to be controversial due to its relation to a numerical model and it was expected to be problematic as this protocol was focusing on qualitative models. The second is very important as it discusses the limitations of the model, and how the DAs evaluate it.

Criterion 8 (Structure-verification) (average 3.8 and mode 4) this is one of the main indicators of the model's presumed utility as it implies the users double-checked the model structure and it is both intelligible and reflects the issues they perceive in the system. Structure verification is present in a major part of the literature dedicated to validation (e.g., Barlas, 1989, 1996; Cassidy et al., 2021; Crielaard et al., 2022; Forrester & Senge, 1980; Lane, 1995; Schwaninger & Groesser, 2020; Sterman, 2000), not to be exhaustive. The reason for that is if the structure is not meaningful for the users if they do not trust it, the model will not be used. The twelve comments on this criterion were very positive showing the users are very satisfied with the structure they produced for the system, even with the limits of representativeness (discussed in criteria 4 and 5). Recommendation: the users must have a clear previous understanding of the difference between causation and correlation and why they are not using correlation to create the structure of the model.

Criterion 9 (Loop Polarity) (average 3.6 and mode 0). The loop polarity identification exercise was not done with the modelers as the strategy from the project is to understand the loops concerning the PESTLE elements identified by the modelers in a future effort, possibly a workshop. Therefore, the thirteen comments received in this criterion reflected the necessity of a session to name and attribute the polarity of the loops (as suggested by Sterman, 2000). Recommendation: understanding the loops provides invaluable knowledge to manage the system and to understand the pathways to manage the variables of interest.

Criterion 10 (Boundary adequacy (as structure)) (average 4.1 and mode 4).

This indicator is very important as it reflects the level of detail vs aggregation used by the modelers to represent their system. The relevant variables must be stated clearly (reducing the noise regarding the important parts) (Balci, 1994; Crielaard et al., 2022;

Forrester & Senge, 1980; Lane, 1995; Meadows, 1980; Schwaninger & Groesser, 2020). The trap is in masking important parts of the system as aggregate variables, making management difficult, and possibly leading to a misunderstanding of the relevant parts influencing the variables of interest. The thirteen comments here were very positive, meaning the level of aggregation was satisfactorily done by the modelers. Recommendation: as simplicity is a goal, discuss with the modelers if a group of variables can be aggregated (keeping its meaning) or if one variable needs to be disaggregated into two or more meaningful variables.

Criterion 11 (Family-member) (average 3.9, mode 4) This indicator tries to embrace the capacity of the model, with few adjustments, to represent a general case, instead of just the case it was built for (Forrester & Senge, 1980; Schwaninger & Groesser, 2020; Sterman, 2000). This indicator can be important to reveal the potential of the models to be scaled up to other regions with similar problems (this was one premise of the project where this protocol was created). Most of the nine comments point out that these models can be used in different regions with similar problems with a small adjustment. Recommendation: building the model in groups increases its level of generality, promoting its usefulness to a general case.

Criterion 12 (Extreme-conditions) (average 3, mode 0). As this variable tries to capture some possible extreme behavior (usually associated with numerical models)(Qudrat-Ullah & Seong, 2010; Sterman, 2000), its applicability in the present case was controversial. The thirteen comments here were confusing as this topic is mostly related to a numerical analysis. We brought it here in case some extrapolations could be used to test the meaningfulness of the structure. Recommendation: if the modelers are not using a numerical simulation, this criterion can be removed to reduce confusion.

Dimension 3: Policy insights and spillovers

This dimension embraces indicators 13 to 18 (Table 1) and has 62 comments. The average result varied from 2.9 (insight generation capacity) to 3.8 (learning). This section of the protocol was dedicated to evaluating the possible messages the modelers already learned that could be used in management (as in Lane, 1995) and how these policies would fit into other systems (Forrester & Senge, 1980; Lane, 1995; Sterman, 2000). This dimension's importance is expected to grow with the maturity of the modeling process. Here the number of “not apply” was large (48%) possibly indicating that several criteria are too advanced for the status of the model.

Criterion 13 (Insight generation capacity) (average 2.9, mode 3) In the present case, the twelve evaluations showed a tie regarding the insight generation capacity. Low insight generation was expected at this initial stage of the modeling process, yet two-thirds of the DAs did produce insights (Appendix D, question B). This indicator represents one of the main final uses of the model (Forrester & Senge, 1980; Lane, 1995), in parallel with the main objective of the modeling process - which is learning (Sterman, 2000) - but depending on the maturity level of the model, it might not be ready to produce these insights. Results showed that even at an early stage, some policy insights can be generated. Recommendation: even in early-stage models, ask the modelers what the main messages the model is passing are and what it implies for policymaking.

Criterion 14 (Relevance and Fertility of PIoR) (average 3.1, mode 0) tries to understand the usefulness of the model in bringing new recommendations and policy insights to manage the system (Lane, 1995). The nine answers here showed the modelers do not identify the recommendations as being relevant, even pointing to the premature stage of the model to provide relevant policy recommendations. Recommendation: if any policy recommendation was provided, ask the modelers to evaluate its relevance.

Criterion 15 (Congruence of PIoR with culture) (average 3.6 mode 0). As this indicator embraces the possible acceptability of the policy insights by the community (Checkland & Poulter, 2020; Sterman, 2000), it would make more sense to apply it when the models are presented to a wider audience. The answers here showed this premature stage, but some respondents pointed out that the models can be potentially acceptable to the broader audience. Recommendation: to evaluate the acceptability of the insights and recommendations for a broader audience, it can be useful to have some workshops, or consultation processes that expose the ideas presented by these recommendations, always considering the variations of people´s views, and culture (Checkland & Poulter, 2020; Oliveira et al., 2024).

Criterion 16 (Boundary adequacy (as policy)) (average 3, mode 3) understanding the boundary adequacy as policy (Forrester & Senge, 1980; Sterman, 2000) reveals if the modelers understand the limitations of their insights but also its potential to seed other similar cases. The majority of the twelve comments here showed the model is a bit premature to extrapolate recommendations, but also some comments reveal that the policy insights produced here can potentially be exported but require some adaptation to the specific cases it will be used. These comments are coherent with those in criterion 11 (family member) as some adaptations are necessary to the model to fit into

another case. Recommendation: explore the spectrum of policy insights concerning the user's culture and with broader goals of equality, justice, and future global scenarios as these goals are coherent with social-ecological systems sustainability.

Criterion 17 (Learning) (average 3.8 and mode 0). In this indicator, there was a confusion of the word stakeholders in its description (as in criterion 7) which made the evaluation confused. Respondents who considered stakeholders as third-party stakeholders answered NA as this model was not shown to third-party stakeholders yet. The other modelers, who understood they were the focus of this question, answered it very positively, showing they learned things about their system they were not aware of, even considering the similarities between the modeler's backgrounds. As learning is probably the most important feature of the process (Sterman, 2000), this indicator brings great feedback about the model and its process and can provide insights to the project coordinator regarding the usefulness of the modeling process. Recommendation: clearly state who are the foci of this indicator and help the modelers to provide feedback on what they learned.

Criterion 18 (Engagement) (average 3.5 and mode 0). This indicator tried to capture if stakeholders already started to change their system after/during the learning they had in this process. Considering the model was built by groups of scientists and not shown to their stakeholders, the results pointed out that no external engagement was perceived so far. The idea in this indicator was to capture if modelers report a change in understanding of the topics made by the stakeholders, to demonstrate this change goes beyond the individual toward a wider group and happened as a spillover from the social network involved in the process, or in other words, if the process promoted social learning (Reed et al., 2010). Recommendation: extrapolations of the knowledge in the modeling process can be done formally (through workshops or university classes) or informally (using personal networks of the modelers), and therefore promoting these spillovers increases the reach of the knowledge and policy recommendations through network connections.

Dimension 4: Administrative, review, and overview

This dimension embraces indicators 19 to 26 (Table 1) and has 87 comments. The average values vary from 2.5 (Replicability) to 3.9 (time and cost of intervention). The spirit of this dimension was to provide insights related to the documentation and replicability of the process, the effectiveness regarding time and cost constraints, and the

capacity of the model to be updated. Additionally, it also considers the possibility of a third-party verification of the model structure and concludes with a verification of real system changes due to the modeling exercise.

Criterion 19 (Ease of Enrichment) (average 3.4, mode 4). This indicator embraces the capacity of the model to be enriched by new data or other assumptions (Lane, 1995). Most of the eleventh answers say it is easy to be enriched and complemented when other information is available. Enrichment is important as learning (Sterman, 2000) brings different knowledge that should be tested and embraced in the model. Recommendation: using tools with low technological requirements (e.g., Vensim) promotes an easiness of enrichment when compared with a more complicated set of tools.

Criterion 20 (Time & Cost of the Intervention) (average 3.9, mode 4). Time and costs are always relevant (Checkland & Poulter, 2020; Lane, 1995; Sterman, 2000) and the majority of the thirteen comments agreed this was a positive aspect of the present case. This compliance of time with the expectations happened because, in modelers' views, some material was sent in advance allowing them to learn about the process. Recommendation: use more interactive tools during workshops (as this might conflict with the simplicity recommended in criterion 19, it is up to the facilitator's judgment what the optimum level of simplicity is).

Criterion 21 (Documentation) (average 3.4, mode 3). This indicator was focused on the capacity to register the steps taken and the data consulted in such a way someone in the future could re-access it (Lane, 1995; Sterman, 2000). The eleven comments pointed out that the process is well documented, but the content of the model is not as it came from workshops. Recommendation: if agreed by all participants, recording the workshops might provide some material for documentation of the content. We see this practice with caution as it might be coercive to people with different views (Checkland & Poulter, 2020) to state their perceptions, once it is being recorded.

Criterion 22 (Replicability) (average 2.6, mode 3) the replicability of the content of a group model building like this is very difficult (Sterman, 2000; Van den Belt, 2004) but the methods to conduct the sessions are replicable and were described in the cited literature. The comments were coherent with this perspective and revealed that more effort in replicability is required. Recommendations are to create a guide or manual for documenting the experience or to have another facilitator write every step

in the model exercise.

Criterion 23 (audit or cross-validation) (average 3.4, mode 0). This indicator tried to embrace a new look at the model, testing if its assumptions and structure make sense from a third-party view. The majority of the eleven comments here pointed out that as the modelers were involved in the process, they are out of the scope of this question (therefore mode 0). Nonetheless, comments also show that the model respects physical laws and does not bring any absurd against established social norms. But we echo the view (Checkland & Poulter, 2020) that conflicting views produce different models, and therefore the real value of this indicator can only be achieved if these different views are considered. That means the model can be checked by third parties but only to a certain extent, as the content of the model will vary according to different views of the modelers against the third parties. Recommendation: clearly state to the third-party review to be strict when regarding physical laws but flexible when comparing with established social norms, culture, or views.

Criterion 24 (higher-level model review) (average 3.6, mode 0). This indicator was dedicated to promoting participation in an external review of the model process, which was not done so far. The comments were provided by the same group who created the model, whose evaluations reflect this view. Recommendation: apply this protocol to the higher-level reviewer or coordinator of the project.

Criterion 25 (Walkthroughs) (average 4, mode 0) This indicator was dedicated to an external group review of the documentation, which has not been formally done. The results were provided by the same group who created the model. This indicator can make sense in a later stage of development. Recommendation: any people who want to provide feedback on the model can be allowed to make a walkthrough.

Criterion 26 (System Improvement) (average 0, mode 0). This indicator represents a utopian expectation that the modelers perceive changes in the system improved by using the outputs of the current study. It is acknowledged the low probability of having such results in early-stage models (Sterman, 2000), but as the performance in this indicator can improve in a later-stage model, and for consistency, this criterion was kept in the protocol. Comments here show it is premature for the present case, as expected. Recommendation: keep the possibility of having some practical results open, even if in early-stage models.