A framework for characterising and evaluating the effectiveness of environmental modelling

Hamilton, S. H., Fu, B., Guillaume, J. H. A., Badham, J., Elsawah, S., Gober, P., Hunt, R. J., Iwanaga, T., Jakeman, A. J., Ames, D. P., Curtis, A., Hill, M. C., Pierce, S. A., & Zare, F. (2019). A framework for characterising and evaluating the effectiveness of environmental modelling. *Environmental Modelling and Software*, *118*, 83-98. https://doi.org/10.1016/j.envsoft.2019.04.008

Published in:
Environmental Modelling and Software

Document Version:
Peer reviewed version

Accepted for publication in Environmental Modelling & Software
https://doi.org/10.1016/j.envsoft.2019.04.008

A framework for characterising and evaluating the effectiveness of environmental modelling

Serena H. Hamilton[1,2], Baihua Fu[2], Joseph H.A. Guillaume[2,3], Jennifer Badham[4], Sondoss Elsawah[2,5], Patricia Gober[6], Randall J. Hunt[7], Takuya Iwanaga[2], Anthony J. Jakeman[2], Daniel P. Ames[8], Allan Curtis[9], Mary C. Hill[10], Suzanne A. Pierce[11], Fateme Zare[2]

[1] School of Science, Edith Cowan University, Joondalup, WA, Australia
[2] Fenner School of Environment & Society, Australian National University, Australia
[3] Water and Development Research Group, Aalto University, Finland
[4] Centre for Public Health, Queens University Belfast, Belfast, United Kingdom
[5] Capability Systems Centre, School of Electrical Engineering and Information Technology, University of New South Wales, Australian Defence Force Academy, Canberra ACT, Australia
[6] School of Geographical Sciences and Urban Planning, Arizona State University, Tempe AZ, USA
[7] United States Geological Survey, Upper Midwest Water Science Center, Middleton WI, USA
[8] Civil and Environmental Engineering Department, Brigham Young University, Utah, USA
[9] Graham Centre for Agricultural Innovation, Charles Sturt University, Wagga Wagga, NSW, Australia
[10] Department of Geology, University of Kansas, USA
[11] Environmental Science Institute, Jackson School of Geosciences, University of Texas at Austin, USA

Abstract

Environmental modelling is transitioning from the traditional paradigm that focuses on the model and its quantitative performance to a more holistic paradigm that recognises successful model-based outcomes are closely tied to undertaking modelling as a social process, not just as a technical procedure. This paper redefines evaluation as a multi-dimensional and multi-perspective concept, and proposes a more complete framework for identifying and measuring the effectiveness of modelling that serves the new paradigm. Under this framework, evaluation considers a broader set of success criteria, and emphasises the importance of contextual factors in determining the relevance and outcome of the criteria. These evaluation criteria are grouped into eight categories: project efficiency, model accessibility, credibility, saliency, legitimacy, satisfaction, application, and impact. Evaluation should be part of an iterative and adaptive process that attempts to improve model-based outcomes and foster pathways to better futures.

Keywords: model evaluation, model assessment, model performance

Highlights

- An evaluation framework for the new process-oriented paradigm of modelling is presented
- Effectiveness of modelling is a multi-dimensional and multi-perspective concept
- 32 criteria for model evaluation are considered from project-level to system-level outcomes
- We link the success of modelling to modelling context and modelling practices

Cite as: Hamilton, S.H., Fu, B., Guillaume, J.H.A., Badham, J., Elsawah, S., Gober, P., Hunt, R.J., Iwanaga, T., Jakeman, A.J., Ames, D.P., Curtis, A., Hill, M.C., Pierce, S.A., Zare, F., (2019) A framework for characterising and evaluating the effectiveness of environmental modelling, Environmental Modelling and Software 118, 83-98.

1. Introduction

In the environmental sector, modelling serves a variety of interrelated purposes, including decision support, scientific discovery, and social learning (Badham et al., 2019; Gober 2018). In the water sector, for example, it supports a range of water management decisions, including infrastructure construction and operations, flood control and drought management, harvesting and storing water above and below ground, maintaining healthy ecosystems, and allocation of water for agriculture, energy production, cities, and environmental uses (Loucks, et al. 2005; Mulligan and Ahlfeld 2016; Snow et al., 2016; Sharvelle et al., 2017; Robert et al. 2018). Modelling also enables scientific discovery, for example, about anticipated impacts of climate change on regional hydrological systems (Cook et al., 2015). It can also serve as a vehicle to accrue and share knowledge in a learning process (Elsawah et al., 2015), and build public interest in adaptive management and collective action (Pahl-Wostl et al., 2010). These modelling functions are increasingly interrelated, stemming from the fact that many environmental, especially water-related, problems are recognised as *wicked* (see Rittel and Webber 1973); in that they are complex, intractable, contentious, and open-ended (Head, 2010).

Evaluation of modelling projects helps improve the modelling when conducted within an ongoing model development process such as that described in phases and steps by Badham et al. (2019) for integrated water resource management. It enables weaknesses to be identified and resolved and can also allow the experience from one modelling project to improve future modelling. Evaluation of a modelling exercise or project is especially important during an era of climate change, globalisation, increasing environmental degradation and high uncertainty in general about the future, requiring effective adaptation in response to changing circumstances (Gorddard et al. 2016; Radhakrishnan et al. 2018). Ecologists now talk about transitioning ecosystems away from conservation to managing the “new normal” (Stein et al., 2013). This transition involves enhanced awareness of system change (e.g. water supply, demand, quality, reliability), and increased interest in the human dimensions of evolving natural systems.

Success in model building and application for challenging interdisciplinary issues is about more than getting the science and engineering right. It is about embedding model building in a social process that links and engages scientists, decision makers, interest groups and the wider public towards achieving impact beyond merely technical performance of a model (e.g. as addressed by Bennett et al., 2013). Such impact can be as basic as sharing understanding of a problem, and as complex as identifying policy changes that yield long-term improvements for society (Ticehurst et al. 2011).

1.1 Towards a new paradigm of modelling and evaluation

In the traditional paradigm of modelling and evaluation, the focus is on the model itself (Pianosi et al. 2016). Knowledge transfer to action is treated as a linear process whereby a model is developed, used, and then subsequently has organizational or societal impact; scientific discovery, decision-making, and social learning are treated as rather separate activities. Over the last decade or so, there has been a transition towards a new process-oriented paradigm where

those three activities are inextricably intertwined (Voinov et al. 2018; Gober 2018). This is a paradigm that we, the authors of this present paper, have been working extensively within. Under the new paradigm, effectiveness of modelling also includes the ability to link scientific discovery, decision-making, and social learning in the modelling process. Modelling outcomes are recognised as being highly dependent on the interaction users have with a model and the modelling process, in addition to the properties of the model output itself. Outcomes therefore emerge in a highly iterative and nonlinear process (Ward et al., 2009). The shift to the new paradigm demands that evaluation not only extends beyond assessment of the model but is also an iterative yet systematic process nested within the social process of modelling.

1.2 Contribution of this paper

This paper intends to expand evaluation of modelling to reflect our transition from the old modelling paradigm where focus is on the technicalities of the model itself, toward the new paradigm where modelling is a social process that considers more holistic outcomes. This work arose from an NSF-funded SESYNC pursuit[1] to integrate understanding of core modelling practices. The framework and concepts were developed as part of a workshop process, based on participants' understanding and supported by literature review. The participants have a diverse range of backgrounds covering social and natural sciences, public health, and computer science, and have extensive experience in the development of models for decision and policy support, social learning and scientific research. This work reflects an iterative process of consensus building rather than a systematic review.

In Section 2, we synthesise existing evaluation literature, and argue that current evaluation practices are largely inadequate in both depth and scope, which limits their applicability and prospects to improve modelling practices. We outline an evaluation process (Section 3) that goes beyond simply confirming a project's achievements to exploring factors and practices that contributed to its success (or failure), providing constructive learning that can feedback into current or future projects. The overall message is that effectiveness comes down to modelling being a process of change rather than change as an outcome. We therefore need to think of evaluation as a nested process in which specific practices are woven together to progressively improve understanding.

Evaluation involves mobilising evaluation criteria that suit the project and evaluation context. The context to be considered is described in Section 4. In Section 5, we expand and present 32 criteria to consider in evaluating modelling, ranging from project-level to system-level outcomes, and from technical metrics to indicators of more complex attitudinal, behavioural and relational changes. This is followed by a brief overview of the common methods used for evaluation in Section 6. The primary contribution of this paper is the overarching evaluation process and the comprehensive list of evaluation criteria. In providing an overview of criteria and techniques and how they fit together, the paper intends

[1] "Pursuits" project at the National Socio-Environmental Synthesis Center (SESYNC) (https://www.sesync.org/for-you/educator/research/themes-pursuits), funded by the United States National Science Foundation (NSF)

to help practitioners make sense of available tools and how they could mobilise these tools for their own needs.

2. Evaluation barriers, gaps and opportunities

There are several objectives in conducting a model evaluation. In terms of accountability, evaluation can provide the research and modelling team, as well as funders and other interested parties, feedback on whether the project is achieving or has achieved its set goals. It can also help demonstrate outcomes from the work. From a policy point of view, an evaluation will help gauge the merit of the work and assist policy makers in determining how much weight to give project/model outputs when making their decisions. An evaluation can also help justify expenditure to funders and provide a guide for resource allocation in the future, including whether funding should be continued, increased or limited. Evaluating the success of projects is also important for learning, sharing and improving the accrual of knowledge. Despite these potentially large benefits of evaluation, a systematic approach to it is not commonly sought.

2.1 Barriers to evaluation

There are several barriers to be surmounted for evaluation to become more commonplace. One is the lack of time and resources to conduct them. Generally, projects are funded up to the point of the delivery of the final model and report, and, in some cases, training in the use of the model. Final project evaluation is rarely budgeted for, thereby creating a lack of incentive to do so (Alexandrov et al. 2011; Schwanitz 2013).

In some cases, the reason for not evaluating success may be structural pressures and biases that modellers are subject to. For models that have been well received, there may be apprehension that further evaluation may uncover shortcomings that undermine the project's performance. On the other hand, for those models that were not well received, there may be reluctance in further scrutinising the work. These are not necessarily conscious motivations for avoiding evaluations, and may occur despite the best intentions of modellers.

Another key reason for not evaluating success may be the lack of awareness or recognition of its benefits, including the view that modelling evaluation is limited to model validation and verification. Through this narrow validation/verification lens, evaluation of complex models such as large integrated socioenvironmental models (Kelly et al., 2013; Hamilton et al., 2015) can be seen as impractical (Jakeman et al. 2006; Schwanitz 2013). This paper expands the definition and conceptualisation of evaluation, enabling modellers to see that even if certain aspects of evaluation cannot be performed (e.g. due to lack of data), others may be both practical and useful.

Other barriers to evaluation may include the limited availability of expertise in evaluation, and the lack of guidance or standard procedures in interpreting and carrying out these evaluations (Alexandrov et al. 2011). This paper intends to help researchers and practitioners overcome these two barriers by providing a framework that guides the characterisation and evaluation of the success of modelling projects.

2.2 Overview of prior work around evaluating success

Prior work has included useful categorisations and criteria for success in several settings, ranging from policy and systems analysis (Goeller, 1988), project management (Ika, 2009; Westerveld, 2003), timescales (Roughley, 2009), decision support systems (McIntosh et al., 2011; Merritt et al., 2017), environmental management and policy (Cash et al., 2003; White et al., 2010), stakeholder equity and representation (van Voorn et al., 2016), and model performance (Bennett et al., 2013). Of note is the landmark paper by Cash et al. (2003) which argued that science and technology are unlikely to be used by decision makers to address environmental problems unless relevant stakeholders see them as credible, salient and legitimate. Discourse has also addressed the value of non-quantitative outcomes such as community and capacity building, and co-learning (Krueger et al., 2012; Voinov and Bousquet, 2010). This existing body of literature defines a diverse set of prior criteria yet each article views success factors through a different lens, resulting in the need to synthesise the criteria and assure that a representative set is defined and described cohesively. These criteria are further discussed in Section 5. Here we give a brief overview of some key lines of work: model evaluation frameworks, factors contributing to project success, and evaluation in environmental planning and participatory research.

There are some available frameworks for evaluating environmental models, however these frameworks are highly technical with a focus on characterising the uncertainty and performance of the actual model (Galelli et al., 2014; Matott et al., 2009; Refsgaard et al., 2007). These frameworks are not suited to evaluating the broader modelling process, including the knowledge building and use processes. At the other end of the spectrum, generic frameworks for evaluating the success of studies in environmental management can be broadly applicable for assessing the social aspects of modelling (e.g. Cash et al., 2003; Goeller, 1988; Roughley, 2009). However, on their own, these generic frameworks capture only a limited depth and/or scope of criteria and considerations relevant to environmental modelling processes. There is a need for a modelling-focused evaluation framework that builds on both existing environmental model frameworks and more generic evaluation frameworks, and provides adequate depth and scope to allow the full range of modelling practices and outcomes to be considered.

The current literature on factors contributing to success is similarly often undertaken within a specific scope, such as for a type of model or tool, or models for a specific purpose. Examples for the first group are abundant, mainly stemming from synthesis of best/good practices for different types of modelling, such as system dynamics modelling (Elsawah et al., 2017a; Martinez-Moyano and Richardson, 2013), environmental decision support systems (McIntosh et al., 2011; Merritt et al., 2017), environmental modelling (EPA, 2009; Jakeman et al., 2006), process modelling (Bandara, 2007), and Bayesian Network modelling (Chen and Pollino, 2012). The second group of studies is less common where examples include those focusing on factors in modelling that contribute to societal problem solving (Sterk et al., 2011), or factors that contribute to a particular evaluation criteria (van Voorn et al., 2016).

The first group, being well-confined and tailored to the processes of a particular type of modelling, may be easier to follow in practice by modellers. In contrast,

factors that are output/outcome-focused seem to be more general and may not be explicit enough for modellers. Also the second group of literature tend to have a clearer and broader definition of success in mind, but in the first group, connection between best/good practices and what constitute “best” or “good” is not always clearly defined. There is a gap in systematically linking modelling context (e.g. purposes, interest groups, resources available), modelling practices and success. Part of this gap is being addressed by studies examining why models and other scientific information have not been used to their full potential (Borowski and Hare, 2007; Diez and McIntosh 2009; Dilling and Lemos, 2011). The major takeaway message from these studies is that greater interaction between users and producers of information improves the usability of results. There is a need to promote evaluation approaches and techniques that are suited to this iterative knowledge exchange process.

There is a considerable amount of literature on evaluation processes in the fields of environmental planning and participatory research. For example, Von Korff et al. (2012) provided an overview of evaluating participatory water management projects. Syme and Sadler (1994) identified six principles for evaluation of stakeholder engagement processes, highlighting the importance of agreeing on objectives of the program and the criteria and methodology of evaluation in partnership with stakeholders, as well as allocating the resources (including evaluators) early in the program. Bellamy et al. (2001) developed an integrated systems-based framework for the evolution of natural resource management policy initiatives, which recognises the multiple levels and nested nature of such policies. Hassenforder et al. (2016b) proposed the Monitoring and Evaluation of Participatory Planning Processes (MEPPP) Framework that includes the consideration of context, process and outputs/outcomes. However, there is little literature that focuses on the evaluation process explicitly for environmental modelling projects.

3. Evaluation process

In the context of environmental modelling under the new paradigm, we suggest that evaluation occurs within an adaptive learning and management cycle, in which evaluation occurs both repeatedly in time, and at different levels of a project (e.g. in designing a stakeholder engagement process, planning a modelling-focussed workshop, and responding to changes dynamically within the workshop itself). This is consistent with the principle identified by Syme and Sadler (1994) that evaluation should influence planning (in our case, projects) on an ongoing basis. To provide feedback to this adaptive cycle, the emphasis of evaluation must shift from being a *summative* assessment to *formative* assessment. While there has been some debate about the distinction between these two types of evaluation (Chen, 1996; Patton, 1996; Scriven, 1991), the key difference lies in their primary function. Summative evaluation passes judgement on whether an aspect of the modelling was effective or not (e.g. ‘was X practice appropriate for realizing Y?’), whereas formative evaluation generates explanatory information about the gap between the actual and desired performance level for the purpose of learning (e.g. ‘why was X practice appropriate/inappropriate for realizing Y?’) (William and Black, 1996). Although both forms of evaluation can be complementary and valuable, more attention is needed on understanding why things went right or wrong (i.e. formative

evaluation) to guide both conceptual and instrumental improvements in ongoing and future projects. Such evaluation cycles can occur within different stages of a project and across different projects.

This cyclic process can be informed by evaluation criteria at various levels of abstraction (Figure 1); these criteria are described in Section 5. The most concrete, detailed, lower level criteria tend to relate to project-level impacts, including project efficiency, credibility, and model accessibility. These criteria are expected to influence higher level criteria including the application of the model and satisfaction of stakeholders. At a higher level, "impact" is a more abstract concept that typically builds on lower level ideas and emerges at system level potentially in the long term. At any point in time, the evaluators ultimately form their own overall judgement of the effectiveness of the modelling project.

It is recognised that bias may exist in the knowledge and perception of the evaluators themselves (Smith et al., 2018). Entirely avoiding bias is impossible, both in modelling and evaluation. However, some techniques applicable to modelling can be used in the evaluation process itself, for example, those that result in a reflective evaluation process whereby in some sense the evaluation is evaluated (see e.g. Lahtinen et al. 2017). This is part of the motivation for the cyclical process presented in Figure 1, which indicates evaluation as an ongoing, iterative process, rather than a once off activity.

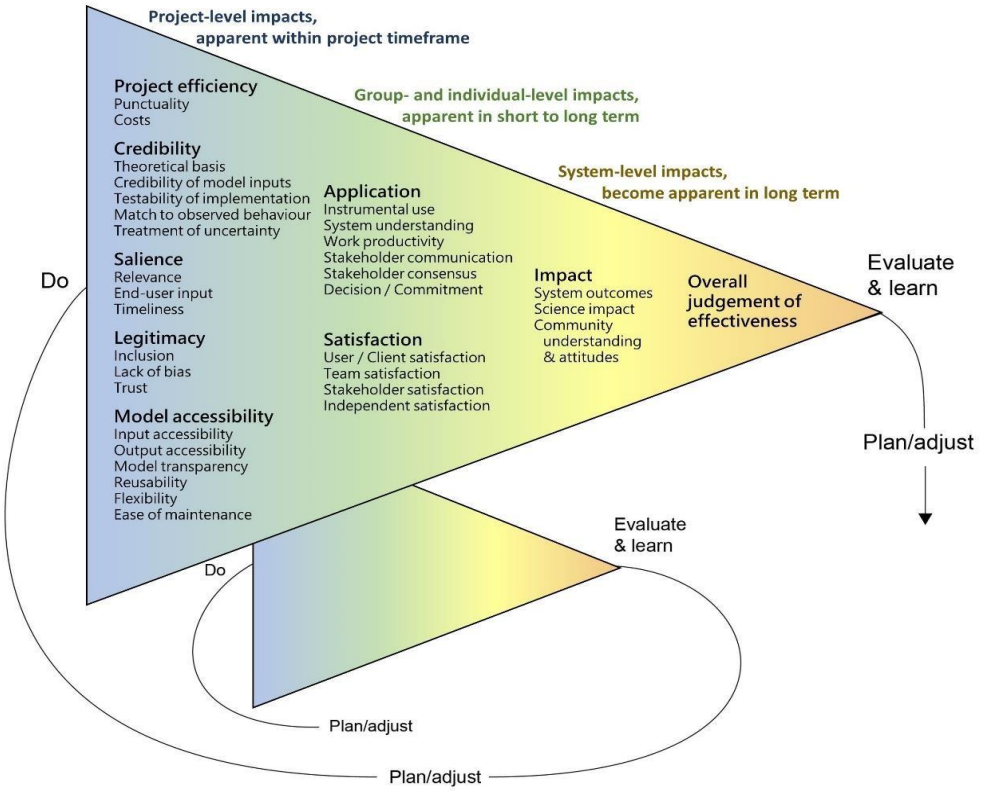

Figure 1 Conceptualisation of evaluation, set within the broader context of an adaptive learning and management cycle, occurring at different scales within a project. Criteria for evaluating effectiveness can roughly be differentiated according to their level of abstraction or detail, with more concrete, detailed, lower level criteria typically influencing more abstract, higher level criteria. Colours indicate level of impact of outcomes, and correspond to circles of influence in Fig. 2.

To describe the evaluation itself, we propose an evaluation process (Table 1 in conjunction with Figure 1) that is applicable to projects across the spectrum from the traditional to the emerging modelling paradigm. It applies both during (*ex ante*) and after (*ex post*) the modelling project and is built on previous research aimed at characterising success of models, decision support tools, and research and management projects in general (e.g. Bennett et al. 2013; Cash et al., 2003; Goeller, 1988; Ika, 2009; McIntosh et al., 2011; Roughley, 2009). Effectiveness can take on different forms, and the relevant criteria depends on the project and evaluation contexts. Overall effectiveness of the project in a sense captures the appropriateness of the modelling approach for achieving the intended objectives, including the selection of tools and methods (Voinov et al. 2018) and their implementation. The form and function of the evaluation activities, described below in Section 4, can vary depending on whether it is an *ex ante* or *ex post* evaluation, and whether the outcomes evaluated are within or beyond the control of the project team. The learnings from the evaluation can then be used to improve current and future projects (Table 1).

Table 1 Evaluation process, involving consideration of context, selection of criteria, execution of the plan and use of evaluation outcomes.

Steps in the evaluation process

1. Identify project context affecting evaluation (Section 4.1)
    - Modelling purpose
    - Problem characteristics
    - Project resources
2. Identify evaluation context affecting method selection (Section 4.2)
    - What scale of outcomes will the evaluation be able to assess?
    - What is the timing and role of the evaluation within the project?
3. Design evaluation process based on the project and evaluation context
    - Select and prioritise evaluation criteria (Section 5)
    - Select methods (Section 6)
4. Execution of evaluation plan and use learnings to improve current and future projects, including adaptive management of the evaluation

4. Context

In applying the evaluation process, it is important to consider the project context, including the modelling purpose, problem characteristics and project constraints (section 3), as well as the evaluation context (section 4.2), which includes the stage of the project at which evaluation is planned, and differing levels of influence over outcomes.

4.1 Project context

Ultimately, the core success of a model is based on the purpose for which it was built (Harmel et al. 2014). At a more general level, model purpose can be considered in terms of three broad categories:

- Decision support tools
- Participatory tools
- Research tools

The purpose of the model strongly influences the relevance of the evaluation criteria used. As a decision support tool, model effectiveness is related to whether it helped improve the ability of an individual or group to make decisions (e.g. Keen 1980). A model can serve as a participatory tool in many forms or degrees of participation (akin to Arnstein's (1969) ladder of citizen participation), ranging from models based on crowdsourced data to models created by participants. As a participatory tool for social learning, effectiveness can relate to how well the modelling supported learning or communication among different parties of stakeholders (e.g. Smajgl and Ward, 2013). Lastly, as a research tool, effectiveness is often centred on whether the modelling helped improve the science and understanding of the system (Duggan, 2015). These three categories are not mutually exclusive, but instead are increasingly intertwined in modern modelling practices. Glynn et al. (2017) argue that such interconnection is needed for adaptive management of natural resources.

The characteristics of the problem can also influence the evaluation criteria, including the benchmark for success. Relevant problem characteristics can include: the number, severity and complexity of issues involved; the diversity of stakeholders involved and their interests or priorities; and the type of system or the system components entailed. For example, there would be a lower expectation of accuracy for a model capturing the system dynamics of a large river basin with multiple and diverse socioeconomic and environmental drivers, compared to that of a model representing a water balance problem for a small, simple, undeveloped catchment.

The resources available to the project, including time, funding, expertise and data, should also be factored in during the evaluation. Such resources can put constraints on what the project can feasibly achieve. It would be unreasonable to expect a project with limited resources to achieve the same outcomes as a project with a large budget and access to a wealth of data.

4.2 Evaluation context

Evaluation design should take into account the relationship between the evaluation and project outcomes, specifically: 1) the scale at which outcomes become apparent for the evaluation; and 2) the role the evaluation plays in the project - how it is to be used or communicated.

4.2.1 *Scale of outcomes and timing of evaluation*

Outcomes of a modelling project can occur at different times and at different levels, influencing the evaluation (Table 2). Outcomes include both tangible outputs and other nontangible benefits of the model and modelling process (see examples in Table 2). Following Rouwette et al. (2002) and Roughley (2009),

we consider outcomes as occurring at the: 1) project level; 2) individual level; 3) group level; or 4) system level, as shown in Figures 1 and 2. Figure 2, adapted from Mendoza et al. (2013), illustrates that the project team has greatest control over the inner circle of the diagram (i.e. project-level outcomes), and the degree of control decreases moving from individual-level outcomes of those directly involved in the project, to group-level outcomes, and finally to the outer circle containing system-level outcomes.

Table 2 Outcome levels and their corresponding outcomes and timeframes

Level of impacts	**Timeframe of outcomes to become apparent**	**Who/what is impacted?**	**Examples of outcomes**
Project	Immediate	• Project products (e.g. models, DSS, tools, methods, findings) • Project methodology	• Completed functional model (validated) • Further use of the products (legacy) • Validation of the methods/findings
Individual	Short to medium/long term	• End-user • Client • Project team members	• Application of the model for its intended purpose • Improved understanding of the system • Improved decision making • Understanding of others' perspectives • Capacity building, skills
Group	Short to long term	□ Stakeholders	• Consensus, trust built • Relationships developed • Exchange of viewpoints • Shared understanding • Joint commitment to action
System	Long term	• Community • Organisation • Environmental asset	• Change in institutional structure/process • Change in attitudes/behaviours • Biophysical changes

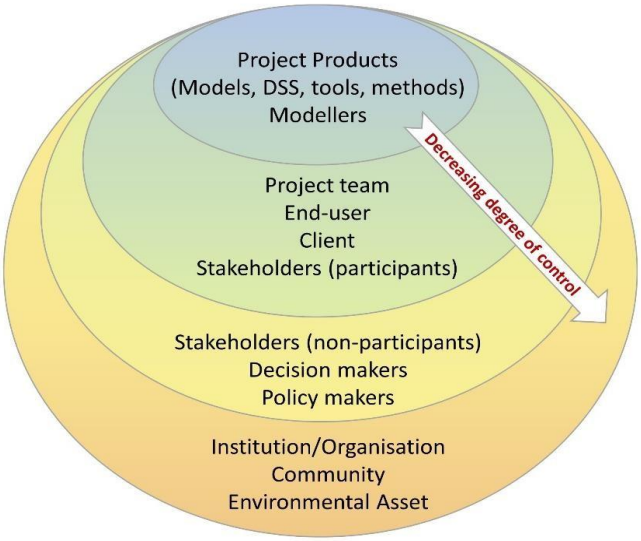

Figure 2 Circles of influence showing who and what is being impacted, and the degree of control the modelling team has over each circle (adapted from Mendoza et al., 2013). Colours correspond to the different levels of impact in Fig. 1.

As outcomes can be time-dependent, the timing of the evaluation can influence its results (Table 2) (Roughley, 2009). For example, an evaluation that occurs immediately after model delivery may find that no system level outcomes were achieved. System level changes may only be detected many years after project completion. Project level outcomes should be apparent within the timeframe of the project itself. Individual and group level outcomes can emerge anytime between the short term and long term. Furthermore, given the behavioural and social nature of most individual and group level outcomes, they are also subject to change through time. The timeframe of outcomes is a function of the project purpose and context. While it can be useful to identify this timeframe so that outcomes are appropriately assessed, it is subject to uncertainty. This uncertainty can be dealt with iteratively within the evaluation process.

Project level outcomes include the delivery of project outputs that are generally agreed on and planned for at the beginning of the project. They may comprise products such as a functional model, or activities such as the provision of training. Project level outcomes may also include criteria related to the quality of the model, for example model validity and the representation of uncertainty.
Outcomes at the individual level refer to the impression or effect of the model on individual end-users, participants or clients. Individual level outcomes may include whether the model was perceived as useful or effective at achieving its intended purpose (e.g. supporting decision making), or whether it provided any new insight or understanding or led to a change of behaviour. It recognises that the same model may have different levels of usage or success on different people (White et al. 2010; Hunka et al. 2013).

Group level outcomes refer to effects of the modelling on the group of participants or stakeholders as a whole. These outcomes, which are particularly relevant for participatory tools, can consist of an exchange or alignment of views, consensus on an issue or solution, building networks, establishing trusting relationships between stakeholders, and increased quality of communication between different stakeholder groups (Forgie and Richardson, 2007; Gray et al., 2017).

System level outcomes relate to changes that occur in the organisation, institution, community or system as a result of the modelling, thereby referring to outcomes beyond the direct control of the model project team, end users and engaged participants. For example, the modelling may prompt changes in how a government agency assesses or manages a resource, including changes to their workflow or processes, or their organisational structure (Halbe et al., 2018). The modelling may have (indirectly) contributed to changes in attitudes and behaviours in the community, for example through someone influential in the community who was exposed to and gained new insight from the model (Diez and McIntosh 2011). System level outcomes also include changes that occur to parts of the system (e.g. the environmental asset or resource) as a result of interventions that were influenced by the modelling.

These system level changes tend to be indirect outcomes, positioned on the outer circle of influence (Figure 2); they are difficult to measure, difficult for the modelling team to control (Roughley, 2009), and therefore the most difficult to evaluate. There are challenges in evaluating system level changes related to difficulties in attributing impact and limits to affecting change. For the former

challenge, establishing causality can be difficult particularly for complex systems, which are subject to a multitude of dynamic processes, and for indirect outcomes. The second challenge refers to limits to the influence that researchbased tools can have on the ground, regardless of the quality of the model and its underlying science (McIntosh et al., 2011). Not only would the model be just one line of evidence, its impact can be limited depending on the politics in play, including the power of individual stakeholders.

For practical reasons, particularly when project resources are limited, the scope of evaluation is typically bound to the inner circles of influence. We advocate an alternative approach to evaluation for system-level outcomes that examines the impact pathway based on the theory of change (Weiss 1995). Using this approach, the team explicitly maps out the impact pathway (i.e. assumptions about the process through which change occurs) toward expected outcomes (Douthwaite et al., 2003). This impact pathway represents the theory of how the team sees the project outputs achieving those system-level outcomes. It helps the team or stakeholders tease out their assumptions and test their validity.

Deliberately thinking about how and why change can happen from the modelling project will help identify factors along each impact pathway that the project team can and cannot influence. This helps the project team to identify and dedicate activities and resources to enable the desired change (Schuetz et al., 2017). For example, capacity building, and maintenance and support of models may be intermediate steps along the impact pathway that are within the project team's influence. The theory of change approach can help identify 'intermediate outcomes', which are indicators to track progress towards the desired outcomes (Douthwaite et al. 2003).

4.2.2 *Role of evaluation within a project*

As discussed early in the paper, there can be different motivations for conducting an evaluation including: exploring project impacts; deepening understanding of the system; improving modelling and other methodological practices; improving and sharing knowledge; assessing merit of the work; and providing transparency by justifying expenditure. Furthermore, the evaluation can be intended for different parties, including the project team (i.e. selfreflection), the funders, stakeholders, and/or other researchers and practitioners. The purpose of the evaluation, including its motivation and who it is intended for, will help determine whose point of view (see Hassenforder et al. 2016b) the evaluators need to consider when measuring the individual criteria.

Ideally evaluation should be considered from the planning stage of the project, including budgeting adequate time and resources to carry out evaluation activities throughout the project as well as post-project. Considering evaluation and identifying criteria from the beginning also helps to provide better clarity to both the modeller and client about what the modelling is trying to achieve. The contextual factors, such as modelling purpose, problem characteristics, and project constraints (Section 3) determine the relative importance of various criteria (described in Section 5).

In addition to the evaluation purpose, evaluation design is also guided by the stage of the project at which it will be performed, and how the evaluation can be used within the project. The use of evaluation results relies on their interpretation

and communication. Interpreting results is a process of linking facts or points collected through data analyses to the purposes and questions that drove the evaluation. Through this process, information turns into evidence that demonstrates progress and success of the project, as well as learnings and suggestions for future improvements. The use of evaluation results may be either summative or formative. In a summative evaluation, one may summarise what aspects of each individual criteria have been done well or poorly. We may also wish to draw overall conclusions on what types of success criteria the project has achieved, and where it may have fallen short. In a formative evaluation, emphasis is on explanation, in which 'why' questions will be asked. For example, if the results are positive and confirm project achievements, what external factors and/or practices were undertaken by the modelling team that contributed to the success? If the results are negative and contradict the planned objectives, what were the factors that contributed to the failure and what could have been done differently?

Evaluation activities conducted while the project is ongoing help assess progress towards objectives and provide feedback to improve the methodology or practices if required. Evaluation of all criteria should be applicable both *ex ante* and *ex post*. However given the timing of *ex ante* evaluations they tend to seek metrics that are only indicative of progress and anticipated outcomes given how the project is tracking. To serve a formative function, the evaluation should not only provide evidence of a gap between actual and desired performance levels for that point of time, but also identify ways to help close the gap (William and Black, 1996). Thus with ongoing evaluation the modelling is carried out in an adaptive learning cycle: if progress is not tracking towards desired outcomes, practices are adapted accordingly. Evaluation at regular intervals of the modelling process can help identify potential issues as they emerge. This enables corrective action to be applied before it becomes a more serious problem requiring major amendments later in the project (Warren, 2014). If evaluation activities only commence later in the project, there is less opportunity for evaluation to improve and influence the project outcomes.

While evaluations conducted *ex post* cannot improve the project itself, they can provide valuable information to improve future projects. With the project complete, many outcomes are final (e.g. project-level outcomes). However, other outcomes may take several years after the project to become apparent (e.g. some group-level outcomes, and many system-level outcomes) and others may vary with time (e.g. related to the use or application of the model). Evaluation may feed into mechanisms to help design future projects.

The role of the evaluation may also evolve during a project. It is typically useful to at least informally reflect on or evaluate the evaluation – checking for bias, whether criteria were realistic/attainable, revisiting the timing of the evaluation and who carried it out. The credibility of the information used as evidence may also be revisited. For example, Bark et al. (2016) evaluated a large integrated project aimed at assessing ecological and economic benefits of environmental water in the Murray-Darling Basin in Australia. Several types of information were gathered including anonymous survey, facilitated workshop and bibliometric analysis of publications. This ensures multiple perspectives and lines of evidences are used to support interpretation of the data against evaluation questions.

5. Evaluation criteria

We have collated 32 criteria for evaluating the success of modelling projects (Figure 1) based on existing literature and the collective experience of the authors. The criteria are grouped into eight categories: project efficiency; credibility; salience; legitimacy; model accessibility; application; satisfaction; and impact. The following subsections describe each of the criteria and the contexts where they may be important or not relevant. Our framework intends to provide the full possible scope within which environmental models can be evaluated, however it is not expected that projects will assess all 32 criteria at a high level of detail. Nevertheless we do propose that all criteria are at least discussed when planning an evaluation.

These criteria will vary in importance depending on the project and evaluation contexts (see Section 4), and should be prioritised on a case-by-case basis. Additionally multiple authors recognise that perceived deficiencies in just one of these interrelated criteria can undermine the overall effectiveness or success of the modelling effort (Cash et al., 2003; Kunseler et al., 2015). For example, engaging more stakeholders in the process may lead to greater legitimacy and salience, but it may also decrease the credibility if the science is no longer viewed to be impartial. Achieving success therefore requires a balancing of tradeoffs across criteria (van Voorn et al., 2016). While it is not feasible to capture the vastly different cases of environmental modelling projects, we suggest the following considerations that may be useful when prioritising criteria.

- Recognise project constraints, which necessitates prioritisation or tradeoffs among the criteria. The most common constraints are related to project efficiency, such as time and cost constraints.
- Identify criteria that are critical to the project objectives.
- Identify criteria that enhance the project objective. These criteria may not be critical, but enhance the use and outcomes of the project, including that of the critical criteria.
- Identify criteria that have little relevance to a particular context. For example, stakeholder communication and consensus may not be important for a biophysical model developed for scientific research.

Here we provide three hypothetical examples of prioritising our criteria for three different types of projects: development of decision support systems (DSS), participatory modelling and development of research models for biophysical systems (Table 3). Note that in general the prioritisation of the criteria should be undertaken and agreed on by all parties in the project, with consideration of the project and evaluation contexts. In addition, we should allow the prioritisation to evolve as we enter different stages of a project.

Table 3: Examples of criteria prioritisation for three different types of projects: development of decision support systems (DSS), participatory modelling and development of research models for biophysical systems.

Level	Category	Criteria	DSS			Participatory modelling			Research modelling		
			Critical	Supportive	Not critical	Critical	Supportive	Not critical	Critical	Supportive	Not critical
Project level	Project efficiency	Punctuality	X				X			X	
		Costs	X			X			X		
	Credibility	Theoretical basis		X				X	X		
		Credibility of inputs	X			X			X		
		Testability of implementation	X				X		X		
		Match to observed behaviour		X		X			X		
		Treatment of uncertainty		X			X		X		
	Saliency	Relevance	X			X			X		
		End-user input	X			X					X
		Timeliness		X		X				X	
	Legitimacy	Inclusion		X		X					X
		Lack of bias	X			X			X		
		Trust	X			X				X	
	Model accessibility	Input accessibility	X			X				X	
		Output accessibility	X			X				X	
		Model transparency		X			X		X		
		Reusability			X		X			X	
		Flexibility		X		X					X
		Ease of maintenance	X					X			X
	Application	Instrumental use	X			X			X		
		System understanding	X			X			X		

Group to individual level		Work productivity		X			X			X	
		Stakeholder communication	X			X					X
		Stakeholder consensus		X		X					X
		Decision/ Commitment			X			X			X
	Satisfaction	User satisfaction	X			X				X	
		Team satisfaction		X			X		X		
		Client satisfaction	X				X				X
		Independent satisfaction			X			X	X		
System level	Impact	System outcomes	X				X			X	
		Science impact			X			X	X		
		Community understanding & attitudes		X		X					X

5.1 Project efficiency

Historically, project management was framed around the three constraints of time, cost and quality (Ika, 2009). The understanding of quality has broadened in recent years and is addressed by the remaining 30 success criteria to varying degrees; therefore this section concerns only time and cost. *Punctuality* and *costs* are the key aspects of project efficiency: the delivery of the project without considering the quality of the deliverables. They are also easily measured, and probably the only completely objective success criteria. Efficiency criteria are relatively generic and relevant for almost all modelling projects.

Punctuality concerns whether the work was completed by the project deadline. Typically, there will be several deliverables at different stages of the project (such as a model, draft paper, final recommendation). While all projects have (or should have) target dates for planning purposes, only some projects have external reasons for these target dates. If the main purpose of the model is to support a decision with a set deadline or provide input to another research project, then the timing impacts on the model's effectiveness, not only the efficiency. It may be necessary to reduce the project scope or make other compromises to meet such a deadline, and a model that cannot be applied as intended may be deemed unsuccessful.

Questions of underutilisation or overload should also be considered in the punctuality criteria. Efficient delivery of a project involves resources (including specialist skills) being available at the time they are required. During the evaluation, critical shortages can be identified so that future projects can consider the ordering of activities that require the scarce resources.

Cost criteria concern whether the project was completed within budget. As with punctuality, the evaluation should consider the allocation of costs to different components of the project, identifying where the component cost is more or less than expected and why.

Time, costs and quality can be traded. For example, additional functions may be added to a model as stakeholders become familiar with how models can be used, but such scope changes impact on the timing and cost of the project. Such trading is linked to client satisfaction with the deliverables (Section 5.7) and involvement in the design; see for example extreme programming, or other agile project management approaches (Beck and Andres, 2005).

5.2 Credibility

Given that modelling is the foundation for evaluating the system of interest for decision making, it is fundamental that stakeholders perceive the model and modelling process as technically and scientifically valid. We consider credibility in terms of five criteria: *theoretical basis* of both the modelling process and model, *credibility of model inputs*, *testability of implementation*, *match to observed behaviour*, and adequate *treatment of uncertainty*. Measures of many of these criteria have been to some degree formalised, for example with good modelling practices (e.g., Anderson et al., 2015; Badham et al., 2019) and concepts of evaluating model reliability (Refsgaard et al. 2004; NRC 2012). Broadly speaking, these measures do not provide yes/no answers (NRC 2012); moreover, the evaluator should bear in mind expectations of different users

(Hunka et al. 2013), tradeoffs between them, and the possibility of changes in expectations over time.

Credibility is fundamentally about whether stakeholders trust that the model and results can be used as purported. When minimising evaluation effort, it may be possible to conclude that because the modelling process is scientifically valid and the system sufficiently understood, model outputs are likely to be credible. Scientific validity of the modelling process may be based on acceptance that the modeller has used accepted or justifiable methods (Voinov et al. 2018), including for data and software management, as well as quality assurance. What is considered acceptable may not have agreement among stakeholders, and credibility of the process may be influenced by the modeller's perceived legitimacy (Section 5.4). Determinations of acceptance of methods typically rely on how widespread the methods are used and discussions of the pros and cons of their use, which is adjunct to model transparency (Section 5.5). It is notable that in an integrated assessment modelling process, a high degree of stakeholder participation throughout the modelling process facilitates development of credibility (Aumann 2011), and is likely to influence the role of the considerations listed here.

5.2.1 *Theoretical basis*

Evaluating the theoretical basis of a quantitative model itself involves determining whether concepts, structure and parameterisation schemes are scientifically justified. Because the natural system is unknowably complex, an evaluation of the model basis is to some extent subjective. At a minimum, the conceptual model must form an acceptable approximation of the modelled system. The extent to which the quantitative model incorporates high degrees of conceptual model complexity is driven by the problem being addressed (Jakeman et al., 2006). A theoretical evaluation involves checking all relevant underpinnings used in the model construction, and that model assumptions invoked are justified. Justification may take the form of formal model confirmation (e.g. Refsgaard et al., 2004), comparison to other models in similar settings, or a check that the model fits client or end-user expectations. When justification is needed, it should support that: 1) the use of the assumptions in the specific circumstances the model is applied; 2) the arguments used are sound; and 3) the assumptions used are not biased toward a modelling outcome. If the theoretical basis of a model is highly uncertain, an associated analysis might instead explore the effect of alternative assumptions, parameterisation schemes, structures or conceptualisations (e.g. Bankes, 1993, Clark et al. 2011).

5.2.2 *Credibility of inputs*

Adjunct to evaluating the theoretical aspects, those model inputs selected for modelling must also be credible. The inputs should be representative of the drivers of the system, and be suited for the required model scenarios. This does not necessarily mean they directly reflect what is expected in the field – they may capture hypothetical what-if situations. Finally, the inputs should be technically correct, without omissions or errors that are not acknowledged.

5.2.3 *Testability of implementation*

The implementation of the model needs to be reliable in its running and output, and technically valid. Verification benefits from being able to examine the code (e.g. open-source code), as well as from a hierarchical testing approach starting with unit tests, benchmark problems and analytical solutions (NRC, 2012). If not already verified by others, evidence that the algorithms solve the salient mathematical equations should be provided (e.g. NRC, 2012). When executed, the software outputs must adequately capture the conceptual model and reflect descriptions from the model domain. The most current version of the software should be used to reduce potential artefacts from bugs in the software, and solutions should be replicable and reproducible. Bugs are inevitable in all software, including high quality commercial software (McConnell, 2004; Refsgaard and Henriksen 2004). Therefore in addition to checking inputs and comparing outputs to observed data, internal consistency checks such as assertions and unit testing, can be valuable in detecting bugs and ensuring the quality of the model code (Crout et al. 2008; Homès 2011).

5.2.4 *Match to observed behaviour*

Stakeholder acceptance of the model is commonly decided, in part, by its ability to simulate what was observed within the natural system. Typically, a model that approximates observed system behaviour well has higher acceptance than one that does not, though other aspects may be more important depending on the context (Olsson and Andersson 2006). To increase acceptance of the model for scenario testing, this may also mean that behaviour is evaluated outside of conditions specified for calibration (e.g., for droughts and other extreme conditions) (NRC 2012, Klemes, 1986). Formal history matching performance metrics (Bennett et al. 2013) objectively quantify the degree of fit between observations and the model's simulated equivalent outputs. History matching metrics may be difficult to construct in highly complex problems, where data is scarce/unavailable, and where the model is numerically unstable or when the model runtimes are extremely long. Future behaviour is notably fundamentally unknown for example in global climate change models
(Schwanitz 2013). Jakeman et al. (2006) note that comprehensive evaluation of behaviour is "rarely possible (or perhaps even appropriate) for large, integrated models." Yet even then, stakeholders or experts may be able to identify aspects that are unrealistic, or judge whether differences from observations are tolerable for the modelling purpose. These judgement might be codified as "stylized facts" describing system behaviours that need to be reproduced (Schwanitz, 2013), minimum performance requirements, or fitness for purpose criteria (Haasnoot et al., 2014; Parker, 2009).

5.2.5 *Treatment of uncertainty*

For model outputs generated in prediction or forecasting mode, assessment of past behaviour typically needs to be complemented by quantification of uncertainty in the actual predictions of interest (Guillaume et al. 2016). Addressing uncertainty requires assessment of variation across many model realisations of possible model inputs for a given structure, and in some cases different model conceptualisations and related structures. That is, one model realisation cannot be considered a full representation of consolidated

knowledge (Bankes, 1993; Maier et al., 2016). More broadly, model uncertainty needs to adequately acknowledge alternative paths that could lead to different modelling outcomes (Lahtinen et al. 2017), and adequately (and legitimately) address disagreements as they arise.

What constitutes adequate treatment of uncertainty is highly problem-specific, but it is generally recognised that formal uncertainty evaluation provides indirect benefits such as increasing the depth of analysis (see Guillaume et al., 2017). Uncertainty quantification can provide better understanding of forecast accuracy (see NRC 2012), which provides for more informed evaluations of risk and reliability. Exploration of sources of uncertainty can be a source of innovation and scientific discovery (Brugnach et al. 2008, p.65). For example, model non-uniqueness can be examined by performing identifiability analysis on hypothetical data that might be collected in the future (Doherty and Hunt, 2009), and help evaluate worth of future data collection (e.g. Fienen et al. 2010). Some model problems benefit from expressing confidence intervals around a prediction/forecast, and there are metrics for measuring quality of uncertainty bounds (Laio and Tamea, 2006; Gneiting et al., 2007; Xiong et al., 2009). Likewise, methods are available to test robustness of management actions (Herman et al., 2015), and a multitude of tools and approaches for working with uncertainty (Refsgaard et al. 2007; Matott et al. 2009, van der Sluijs et al. 2005;Jakeman et al. 2006). In specific domains, there may be guidelines or stated protocols for evaluating the treatment of uncertainty, but typically each modelling project will have its own criteria for determining whether adequate consideration is given to whether alternative paths might have resulted in different outcomes.

5.3 Salience

In terms of modelling, salience refers to the potential usefulness of the model and/or modelling process to the end user. Can the model help address users' questions of interest? How likely are decision-makers to integrate results into policy decisions? Salience is critical to all model and evaluation purposes. The actual use and impact of the model is further discussed under "Application" and "Impact". The focus here is on the purpose of the modelling: is the model relevant, did end-users provide input into the design of the modelling, and is it timely? This corresponds to the three criteria: *relevance*, *end-user input*, and *timeliness*.

5.3.1 *Relevance*

Science methods and outputs need to link to local issues relevant to stakeholders. If a model is not relevant to addressing the end-user questions, whether that be related to decision making or scientific research, it cannot be deemed successful. As put by Lusiana et al. (2011): "*perfect models that only answer irrelevant questions in users' perspectives have limited utility.*" Salience relates not only to the inclusion of important input and output variables, but also to their adequate representation, including appropriate spatial and temporal extents and resolutions that capture the variables. This criterion therefore relates closely to the modelling being fit-for-purpose, a central tenet of good modelling practice (Jakeman et al. 2006).

5.3.2 *End-user input*

End-user engagement is often considered critical for ensuring relevance of model results. The USA National Research Council asserted that "*inadequate progress has been made in synthesizing research results, assessing impacts of climate change on human systems, or providing knowledge to support decision making and risk analysis*" (NRC 2007). Reasons included a lack of meaningful interaction between scientists and decision makers, and difficulty in interpreting scientific information and results and translating them into recommendations for action. Dunn and Laing (2017) came to similar conclusions about the disconnection between research and policy more recently. Rather than assuming stakeholder needs are known, it is useful to have users at the discussion table to frame the problem initially, or establishing two-way, iterative engagement between producers and users to build trust and better understand the needs of policy-makers (Dilling and Lemos, 2011).

Furthermore end-user engagement is important from a social learning perspective, which is the idea that people learn in groups; this has emerged as imperative for mediating the science-policy divide (Gober, 2018; Gynne et al. 2017). Managing environmental systems in an era of uncertainty requires the capacity to learn from experience, synthesize different types of knowledge and experiences, and view policies as learning experiments (Folke et al. 2005).

From a water resource perspective, this means greater interaction and mutual learning from scientists and water managers. The need for social learning implies new roles for scientists in the water management process, moving from providers of scientific tools and insights to partners in the use of tools and insights for policy and adaptation decision-making (Pahl-Wostl, 2009; Pahl-Wostl et al. 2012).

Direct end-user engagement may be beneficial, including co-authorship between scientists and decision makers. However, it is not always appropriate for modellers to engage directly with users. Instead, engagement might be structured through knowledge brokers or boundary organisations who can negotiate tensions and facilitate useful exchange between scientists and decision makers, linking science and decision-making, and building collaboration and cooperation (White et al. 2008; Crona and Parker, 2012). Boundary agents help ensure that scientists provide information that fits the given policy context. Boundary objects such as water resource models are an obvious way for the two groups to work together, but uneven power relations and the institutional differences between academic and public sector employment can stymie their role in mediating interests of the two groups (White et al. 2010).

5.3.3 *Timeliness*

Even where modelling is in principle relevant and end-users are appropriately engaged, the timeliness of scientific activities needs to be right (Cash et al. 2003). If modelling exercises are run or model results are published at an inconvenient or inopportune time (practically or politically), then they may be less likely to be salient or have an impact. In a policy context, timing needs to fit in the policy decision window when the need for change is widely acknowledged and participants feel they can make a difference (Huitema and

Meijerink 2010). This is not to say that modelling should not proceed outside those times. Indeed, often modelling exercises need to have occurred prior to the decision window in order for model results to be available. Rather, the purpose and design of the project and its evaluation should be adjusted to reflect these timing constraints. For example, a modelling exercise that loses its salience due to political events can still be effective if it instead aims to ready materials for the next decision window – and the evaluation needs to reflect this new aim.

5.4 Legitimacy

Legitimacy is the extent to which decision makers or stakeholders feel that the science or model was developed and presented in an open and unbiased way, respectful of divergent points of view in the community (Cash et al. 2003). This criteria includes the acceptance of the authority of the modelling process to influence decision making (based on Lockwood et al. 2010). Legitimacy is closely related to notions of fairness and justice (Syme et al. 1999). It is critical for all models, particularly when there is a human dimension involved, which is the case for any model focussed on decision support. Without legitimacy of the model, subsequent scientific advice and decisions may also not be perceived as legitimate. Legitimacy may often be observed retrospectively, through increasing commitment of decision makers, from exploratory conversations and brokering the pros and cons of potential decisions to later events that mobilize action on agreed upon policy. We focus here on early indicators: *inclusion*, *lack of bias* and *trust*.

5.4.1 *Inclusion*

Stakeholders need to feel that the modelling process has included them and/or their perspectives. The presentation of scientific results and model outputs is inherently a political process, and therefore one that must reflect a community's divergent viewpoints. Increased acceptance of resultant management decisions is part of the motivation of engaging stakeholders throughout the process from the early stage of defining the problem and identifying priorities and constraints through to interpreting model results, and involving them in the process as partners (Jakeman et al. 2006; Röckmann et al., 2012). Scenarios are often considered a powerful method of engagement, particularly focussed on alternative visions of the future and outcomes of competing policy decisions (Larsen and Gunnarsson-Östling 2009).

Stakeholders may also judge legitimacy based on the inclusion of others – does the modelling draw on a wide range of opinion and input from stakeholders, including community representatives, minorities, and Indigenous groups. Water decisions, for example, are inherently valued-based, and they reflect the meaning and purpose of water to different groups in society. Underlying values about water stem from beliefs about human rights to water, economic efficiency, social equity, environmental protection, provision for future generations, and the role of government, as well as aesthetic and spiritual concerns. Today's water policy decisions reflect deep-seated beliefs about the rights and responsibilities of individuals and groups in society and the role of science in decision-making

(Gober 2018). Incorporating the range of beliefs about water from a community reflects the legitimacy of the decision-making process in that community.

5.4.2 *Lack of bias*

Stakeholders should be included in an unbiased way. Bias in the model can be perceived if it reflects goals or perspectives held only by one group of stakeholders or preconceptions held by the modeller, or disregards those of another stakeholder group (White et al. 2010). This does not necessarily mean, however, that every perspective needs to be given equal voice. Rather, stakeholders need to be included in a fair way, which may depend on the size of each stakeholder group affected, their stake in the issue, or formal rules about standing or admissibility in a legal setting. Legitimacy rests on the capacity to muster a representative cohort of decision-makers and the public – on their terms.

5.4.3 *Trust*

Legitimacy is fundamentally about accepting others' contributions to decision making, meaning that trust plays an important role. Trust is about the willingness of those in a dependent relationship to rely on each other (Sharp and Curtis 2014). This applies not just between researchers and other stakeholders, but also within a research team. Individuals vary in their predisposition to trust others, but their willingness to rely on others is also based on their assessments of the trustworthiness of others. Trustworthiness is based on assessments of ability, integrity (do they hold/exhibit desirable values), and benevolence (to what extent do they consider my interests) (Mayer et al. 1995). Trust may take time to form, and can be influenced by previous experiences with not only those individuals involved but also their institutions or related groups (e.g. water experts or modellers in general) (Olsson and Andersson 2006). Considerations around trust may be related to trust in a model or those doing the modelling – assessment of legitimacy may wish to consider both.

5.5 Model accessibility

5.5.1 *Input and output accessibility, and model transparency and traceability*

The utility of the model strongly depends on its accessibility in terms of the usability of the model and its outputs and how well they are understood. We consider three criteria relevant to model accessibility in the immediate to short term and another three criteria, described later, relevant in the medium to long term. The immediate to short term criteria are: *input accessibility,* which relates to the ease of use of the model by the intended end user to perform the task for which it was designed, including the effort required to preprocess data as model input; *output accessibility* which relates to whether the model results can be understood, again to the intended audience; and *model transparency* which refers to whether the inner workings of the model are available to users. Model transparency includes the accessibility of the theory and assumptions underpinning the model to enable a deeper interpretation of the model results. Comprehensive documentation of the rationale of the model, its development

process, the intended area of application and its limitations can reduce uncertainty about how the model can be applied (Crout et al. 2008).

Key to these three short-term accessibility criteria is ensuring the model and its outputs are suited to the target audience, whether they be decision makers, scientists or community members. These criteria are relevant to all models, regardless of the project context. The model should be designed to bridge the gap between the technological aspects of the model and the cognitive aspects of end users, shaped by their background and technical levels. This bridging can be achieved, for example, through the design of a user-friendly software interface (GUI) and provision of a non-technical (e.g. written in plain English) user manual for operating the model and interpreting results (McIntosh et al. 2011). If the system is non-intuitive to the end-user and difficult to navigate, long-term adoption (especially with staff turnover) may not be achieved even if training is provided in the beginning. Others have also found that the complexity of models may contribute to model rejection (Kolkman et al. 2016).

On the other hand, it has also been argued that a user interface that is too simplistic can reduce the transparency of the analyses occurring within environmental models, undermining the user's satisfaction that the complexity of the problem has been adequately captured (Matthews et al., 2011; Stirling, 2010). This suggests the importance of matching the model's degree of complexity and transparency with the user's capacity and expectations (Gilbert et al., 2018). For some users to trust the model, it may be important that it not be a black box and users are able to access and trace the logic of its complex inner workings; this may be through documentation.

Ideally, model accessibility should be tested by end users continuously throughout the development process and before the model is delivered (i.e. formative evaluation; user acceptance testing) to allow modifications to the design of the model to better suit end user requirements (Otaduy and Diaz, 2017). These accessibility criteria may not be perceived as important in many evaluation contexts in comparison to more outcome based (e.g. system level) criteria. However, model accessibility is critical to its use and may be the underlying factor determining whether or not a model is actually used and has subsequent impact.

5.5.2 *Reusability, flexibility and ease of maintenance*

The three shorter term criteria discussed above (Section 5.5.1) then interplay in the medium to long term through influencing model *reusability*, *flexibility*, and *ease of maintenance*. The importance of these additional criteria varies depending on the purpose of the model. Reusability refers to both i) running and re-running the model and ii) the action of repurposing the model for other applications and contexts. Flexibility on the other hand refers to the ease with which modifications can be made to include additional processes or exclude less relevant processes to better fit a model for its purpose. Ease of maintenance is defined as an attribute of the model that enables defects or issues to be identified and resolved with minimal effort on the part of the model maintainer. This can be achieved through the use of software testing principals, including the creation and maintenance of a test suite, which aids in alerting the

maintainer when a change to the model causes adverse effects or has unintended consequences (Crout et al. 2008; Homès 2011).

Model reusability hinges on the technical implementation details of the model (how it was developed) and how well its use is documented (Holzworth et al. 2010). A model cannot be considered reusable if it cannot be applied to a similar but new context without significant modification to the underlying code. Similarly, insufficient user documentation hinders model reusability as the model cannot be reapplied to a new context without the user having prior knowledge of how to do so – an example of model transparency affecting reusability.

Flexibility plays an important role in the medium term as poor model flexibility negatively impacts development velocity – the speed at which improvements and modifications can be made. An inflexible model structure hinders the ability to incorporate new information, knowledge, and data, such as those that may come to light through an iterative development process (Krause and Flügel 2005; Formetta et al. 2014). In the longer term, lack of flexibility compromises model reusability as relevant processes may not be adequately captured for the model's intended purpose.

It is not advocated here that all models be made reusable or flexible, and expending considerable energy on ensuring ease of maintenance may not be appropriate. These all depend on the given modelling purpose. However, the benefits of models that may be repurposed and adjusted for different contexts is increasingly acknowledged by the environmental modelling community (de Kok et al., 2015). Approaching model development in this manner increases a model's flexibility of use. Repurposing a research tool for use in participatory or decision support contexts is better achieved if code and processes are documented and changes to the code base do not adversely and unnecessarily propagate throughout the model structure. As a beneficial side-effect, such ancillary support processes increase a model's ease of maintainability. However, adopting development approaches to support reusability and flexibility is often a secondary concern (de Kok et al., 2015; Hutton et al., 2016).

These accessibility criteria may not be a factor in cases where use of the model beyond its initial purpose is not intended. This may be in cases where a single context-specific model is agreed to by end users prior to the delivery of the model. It is notoriously difficult and costly, however, to introduce reusability, flexibility, and ease of maintenance after the fact. Such difficulties are well documented in both domains of software and model development, giving rise to iterative development practices (as evidenced by Jakeman et al., 2006; Larman and Basili, 2003). Where in line with the purpose of model development, the criteria of model reusability, flexibility, and ease of maintenance should ideally be considered from the beginning of the development process.

5.6 Application

This group of criteria concerns the application of the model and the direct impact of its use. Our first two criteria in this group correspond to the components of utilisation success proposed by Goeller (1988): *instrumental use* which is the use of the model by the intended end users for the intended purpose; and

conceptual use which we refer to as *system understanding,* i.e., improved understanding of the system or problem as a result of model use or involvement in the modelling process. The third criterion, *work productivity,* considers whether the model results in improved work effectiveness or efficiency for the end user. If work productivity is not improved, there may be no advantage nor incentive to adopt the model.

The relevance of the other three criteria in this group depends on the model purpose: *stakeholder communication* considers whether the modelling process helped facilitate communication between stakeholders, including an exchange of viewpoints and understanding of other participants; *stakeholder consensus* considers whether the modelling process helped stakeholders (or at least the participants) arrive at a shared view of the problem or actions required; and *decision making* considers whether the modelling process or results influenced actions taken to address the problem, including increased commitment to address the problem.

These application criteria correspond to individual and group level criteria. The first three criteria – instrumental use, system understanding and work productivity – are expected to be relevant to most contexts. How these criteria are interpreted and assessed is highly dependent on the model purpose. The stakeholder communication and consensus, and decision criteria, are relevant to models serving participatory and decision support purposes. It may be possible to assess all five application criteria shortly after the delivery of the model, however some of the criteria may be subject to change over time, so an evaluation of the criteria in the medium to long term may also be appropriate. For example, the end user may take several months or longer before they are comfortable with using the model to its full capacity, or the participatory exercise may have led to initial contact between stakeholders with a notable improvement in communication between the parties occurring long after project completion. Conversely, apparent success in the instrumental use of the model or work productivity may change over time, for example, if the model becomes outdated and is too difficult for end users to update. Therefore other criteria, such as accessibility, may be tied to these application criteria.

These application criteria can also be assessed in terms of the available project resources. For example, the evaluation may consider whether these outcomes could have been achieved more easily or cheaply (i.e. efficiently) using another approach. Such consideration may be particularly relevant if the evaluation purpose is to justify expenditure.

5.7 Satisfaction

The success of a model can also be gauged by the appraisal of the end product and modelling process by the different groups of people associated with the project or the modelled problem, including the end-user or client who funded the project, the project team, stakeholders or independent reviewers. In many ways, project satisfaction is an aggregate measure of all criteria perceived as important by the individual or group. As with any summative judgement, this assessment will be influenced by the respondent's personal attributes (e.g. values, beliefs, personal norms, knowledge and skills) as well as their experience with the model, including their level of engagement with the process,

and their understanding and expectations of the model and/or modelling process (Olsson and Andersson 2006; Hunka et al. 2013).

The purpose of the model as well as the purpose of the evaluation will determine the value of the respective satisfaction criteria, i.e. whose satisfaction is important? For instance, for a research tool, independent satisfaction (e.g. expert peer review) may be most important. For a decision support tool for operational management, end-user satisfaction is fundamental, but a decision support tool for management of more controversial issues, such as water allocation, may require satisfaction by stakeholders such as the affected communities. On the other hand, stakeholders engaged with the participatory modelling process, may provide valuable feedback on the social learning achievements of the project. The satisfaction of the project team may be useful for evaluating the value of methodological practices, but may be considered too biased for assessing the overall merits of the work.

Satisfaction can be assessed at or after completion of the project. If this appraisal is undertaken shortly after the project is delivered, the detailed aspects of the project are more likely to be recalled. However, many outcomes, particularly system-level impacts, may yet be realised. On the other hand, satisfaction appraisals conducted many years after project completion may be able to capture more types of outcomes, as well as information on long-term model usage or changes. However, as time goes on it may be more difficult to engage with the relevant people, for example staff members originally engaged with the project may have left the organisation.

5.8 Impact

This final group of criteria concerns the more system-level outcomes of the model. *Community understanding and attitudes* refers to whether the model helped improve awareness, knowledge or confidence in science, or influenced the attitudes or behaviours of the wider organisation or community. *System outcomes* considers whether the model has led to changes in the problem situation. System outcomes tend to be indirect, such is the case where the modelling or model results influenced a decision that was implemented and had on-the-ground outcomes. Finally, *science impact* refers to whether the modelling generated new insights in the research field; this may include improved understandings in the methodology (e.g. modelling approach), the field of application, or across disciplines.

These criteria tend to be in the outer edge of the circles of influence of the project (Figure 2), and therefore difficult to affect as well as to evaluate. The impact of the model is not only contingent on the application of the model, but also various external factors often beyond the control of the project team (e.g. politics, natural processes, other competing socioeconomic objectives), which give rise to the challenges of attributing impact and limits to affecting change. Moreover, models are typically just one of many lines of evidence used in decision making, and in scientific and participatory contexts. Even very good models may fail to make an impact due to other factors. For instance, a model that leads to an agreement to major reductions in water extractions, may result in no positive outcomes in the environment due to drought conditions. Similarly there are challenges associated with translating system understanding to

changes in behaviour, particularly at a community level (Kollmuss and Agyeman 2002).

The relevance of the criteria are dependent on the purpose of the model. These impact criteria are likely to be best assessed in the longer term, as it is unlikely that outcomes will be achieved immediately. In addressing the difficulties in attributing impact, it may be most appropriate to map out the impact pathway (Douthwaite et al. 2003; see Section 4.2.1) and identify and assess those intermediate outcomes that are somewhat within the project team's influence.

6. Overview of evaluation methods

There is a wide range of methods that can be used to collect data for evaluation (Harvey, 1998; Boaz et al, 2008), which reflects the fact that there is no single best method. Several factors are to be considered when selecting data collection methods, such as: What is the objective of the evaluation? What are the resources (time, cost, skills) available for the evaluation? What are the constraints that may determine the evaluation (e.g. cultural and political conditions)? It is often the case that the researcher will need to combine more than one data collection method in a mixed method approach to complement and compare results from different methods. Table 4 gives an overview of the evaluation methods, along with their resource requirements, strengths and limitations.

7. Conclusion

Our proposed framework embraces a more complete perspective on evaluation, extending beyond an assessment of the model product to consider the entire modelling process. It emphasises the flexibility and interactivity of the social process that takes place when models are developed, applied, and shared with stakeholders. The whole process of model development, application, and communication ultimately affects whether knowledge and models are used for decision-making and/or achieve useful impact in other nontechnical ways.

To serve the new process-oriented paradigm of modelling and evaluation, our framework characterises effectiveness as a multi-dimensional and multiperspective concept covered by 32 criteria. These evaluation criteria range from project-level to system-level outcomes, and from quantitative measures of the technical validity of the model to more complex indicators, such as consideration of if and how the modelling process has affected attitudes and behaviours of end-users or beyond. Each model and project has a unique context and purpose, and as such there cannot be a standard benchmark against which to judge their effectiveness. Rather, for each case, the evaluation methods and criteria are determined by the project and evaluation contexts, including the aims, priorities and constraints of the project and the evaluation purpose and the scale of outcomes of interest. There may be value in building a database of modelling evaluations (e.g. checklist or narratives under each criteria) to facilitate learning across projects especially those with similar contexts.

The evaluation process is described as an iterative process nested within an adaptive learning and management cycle, to promote constructive learning that can feedback into the ongoing project and future endeavours. To improve outcomes it is critical that evaluation be factored into project plans and budgets. Scientific discovery, decision making and social learning have become increasingly intertwined in modelling processes within the field of managing natural resources. Therefore evaluation should be part of an ongoing exchange between producers and users of knowledge, including modellers and decision makers, to improve model-based outcomes and promote pathways towards positive futures.

Table 4: Overview of methods used for evaluation

Methods	What it might include?	Skills	Time and cost	Strengths	Issues to be considered	Examples from literature
Openended interviews	Face to face or over the phone questions with key stakeholders or model users	Interviewing, narrative and discourse analysis	Time demanding for both participants and the evaluator	Open questions and probes allow for indepth data	Open responses are relatively difficult to analyse	Blackstock et al. 2007, Jones et al. (2009); Hassenforder et al (2016a); Perez et al (2014)
Surveys	Systematic questionnaires administrated over mail, phone, online, or in person	Questionnaire design, statistical analysis	Relatively not time demanding for participants and the evaluator	• access to large sample size • anonymous responses • the standardized responses are relatively easy to analyse	• Closed questions do not capture rich data • Sample representation • Low response rate	Jahangirian et al (2018), Merritt et al (2017), Crochemore (2011), Bellocchi et al (2015)
Focus groups	Planned discussions or workshops with a small group moderated by a trained facilitator	Facilitation skills	Relatively less time demanding for both participants and the evaluator	• Collecting data from a group of people at the same time. • Provide insights about group interactions	• Group effects and dynamics • Group composition (e.g. power levels, background)	Matthews et al (2011); Tavella and Franco (2015)
Pre and post testing	Collecting qualitative or quantitative data (through surveys, interviews..etc)	Experimental design, statistical analysis	Time demanding for both participants and the evaluator	Provides comparable datasets to test and measure the effects of interest, and the variables that	Sensitivity to many factors that influence the measurements, such as the time between data	Elsawah et al (2017b), Stave et al (2015); Smajgl and Ward (2015)

	at multiple points of time			influence the generated effects	collection, sample size Logistics of engaging participants multiple times	
Observation	Participant or non-participant, recorded through notes or videos	Ethnographic inquiry skills, such as discourse analysis	• Does not demand time from participants • Time consuming for the evaluator	• In situ learning about the system/process of interest • Can reveal unanticipated insights • Flexible data collection method	Observed behaviour can be difficult to interpret and link to the evaluation objectives	Franco and Greiffenhagen (2018); Stave (2010)
Document analysis	Analysis of policy documents, reports, statistical data, and projects memos	Varies depending on the data type, but basic desktop review and analysis skills are needed	• Does not demand time from participants • Time consuming for the evaluator	Data already exists	Data can be limited or incomplete	Seidl (2015); Hassenforder et al (2016a)
Informal evaluation methods	Informal conversations, meetings	No particular skills	Relatively less time demanding from participants and the evaluator	Useful for the internal use of the research team	Difficult to present formally for an external audience	Jones et al. (2009)

8. Acknowledgements

This work was supported by the SESYNC (Socio-Environmental Synthesis Center) Core Modelling Practices in IWRM project under funding received by the National Science Foundation DBI-1052875. Joseph Guillaume was supported by Academy of Finland funded project WASCO (grant no. 305471) and Emil Aaltonen Foundation funded project 'eat-less-water'. The authors thank four anonymous reviewers and Alexey Voinov for their useful comments.

9. References

Alexandrov, G.A., Ames, D., Bellocchi, G., Bruen, M., Crout, N., Erechtchoukova, M., Hildebrandt, A., Hoffman, F., Jackisch, C., Khaiter, P. and Mannina, G., (2011). Technical assessment and evaluation of environmental models and software: Letter to the Editor. Environmental Modelling & Software, 26(3), 328-336.

Anderson, M., Woessner, W.W., Hunt, R. (2015). Applied Groundwater

Modeling, Second Edition: Simulation of Flow and Advective Transport. Academic Press. https://doi.org/10.1016/B978-0-08-091638-5.00001-8 Arnstein, S.R. (1969) A ladder of citizen participation. Journal of the American Planning Association, 35(4), 216-224.

Aumann, C.A. (2011). Constructing model credibility in the context of policy appraisal. Environ. Model. Softw. 26, 258–265.

https://doi.org/10.1016/j.envsoft.2009.09.006

Bandara, W. (2007). Process modelling success factors and measures. (Doctoral dissertation), Queensland University of Technology.

Badham, J., Elsawah, S., Guillaume, J.H.A., Hamilton, S.H., Hunt, R.J., Jakeman, A.J., Pierce, S.A., Snow, V.O., Babbar-Sebens, M., Fu, B., Gober, P., Hill, M.C., Iwanaga, T., Loucks, D.P., Merritt, W.S., Peckham, S.D., Richmond, A.K., Zare, F., Ames, D., Bammer, G., 2019. Effective modeling for Integrated Water Resource Management: A guide to contextual practices by phases and steps and future opportunities. Environmental Modelling & Software 116, 40-56.

Bankes, S. (1993). Exploratory Modeling for Policy Analysis. Oper. Res. 41, 435–449. https://doi.org/10.2307/171847

Bark, R. H., Kragt, M. E., Robson, B. J. (2016). Evaluating an interdisciplinary research project: Lessons learned for organisations, researchers and funders. International journal of project management, 34(8), 1449-1459.

Beck, K., Andres, C., (2005). Extreme Programming Explained: Embrace Change, 2nd Edition. Addison-Wesley, Boston.

Bellamy, J.A., Walker, D.H., McDonald, G.T. and Syme, G.J., 2001. A systems approach to the evaluation of natural resource management initiatives. Journal of environmental management, 63(4), pp.407-423.

Bellocchi, G., Rivington, M., Matthews, K., & Acutis, M. (2015). Deliberative processes for comprehensive evaluation of agroecological models. A review. Agronomy for Sustainable Development, 35(2), 589-605.

Bennett, N.D., Croke, B.F.W., Guariso, G., Guillaume, J.H.A., Hamilton, S.H., Jakeman, A.J., Marsilli-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. Environmental Modelling and Software 40, 1-20.

Blackstock, K. L., Kelly, G. J., & Horsey, B. L. (2007). Developing and applying a framework to evaluate participatory research for sustainability. Ecological economics, 60(4), 726-742.

Boaz, A., Fitzpatrick, S., & Shaw, B. (2008). Assessing the impact of research on policy: A review of the literature for a project on bridging research and policy through outcome evaluation [R/OL]. London: King's College London, 2008.

Borowski, I., Hare, M. (2007). Exploring the gap between water managers and researchers: difficulties of model-based tools to support practical water management. Water Resources Management, 21, 1049-1074.

Brugnach, M., Pahl-Wostl, C., Lindenschmidt, K.E., Janssen, J.A.E.B., Filatova, T., Mouton, A., Holtz, G., van der Keur, P., Gaber, N., (2008). Chapter Four Complexity and Uncertainty: Rethinking the Modelling Activity, in: Jakeman, A.J., Voinov, A.A., Rizzoli, A.E., Chen, S.H. (Eds.), Environmental Modelling, Software and Decision Support. Elsevier, pp. 49–68. https://doi.org/10.1016/S1574-101X(08)00604-2

Cash, D.W., Clark, W.C., Alcock, F., Dickson, N.M., Eckley, N., Guston, D.H., Jäger, J., Mitchell, R.B., 2003. Knowledge systems for sustainable development. Proceedings of the National Academy of Sciences of the United States of America (PNAS) 100 (14), 8086-8091.

Chen, H.-T. (1996). A comprehensive typology for program evaluation. American Journal of Evaluation, 17, 121-130.

Chen, S. H., Pollino, C. A. (2012). Good practice in Bayesian network modelling. Environmental Modelling & Software, 37, 134-145.

Clark, M.P., Kavetski, D., Fenicia, F., (2011). Pursuing the method of multiple working hypotheses for hydrological modeling. Water Resour. Res. 47. https://doi.org/10.1029/2010WR009827

Cook, B. I., Ault, T. R., Smerdon, J. E. (2015). Unprecedented 21st Century Drought Risk in the American Southwest and Central Plains. Science Advances, 1(1): DOI:10.1126/sciadv.1400082.

Crona, B. I., and J. N. Parker. 2012. Learning in support of governance: theories, methods, and a framework to assess how bridging organizations contribute to adaptive resource governance. Ecology and Society 17(1): 32. http://dx.doi.org/10.5751/ES-04534-170132

Crout, N., Kokkonen, T., Jakeman, A.J., Norton, J.P., Newham, L.T.H., Anderson, R., Assaf, H., Croke, B.F.W., Gaber, N., Gibbons, J., Holzworth, D., Mysiak, J., Reichl, J., Seppelt, R., Wagener, T., Whitfield, P., 2008. Good modelling practice. In: Jakeman, A.J., Voinov, A.A., Rizzoli, A.E., Chen, S.H. (Eds.), Environmental Modelling, Software and Decision Support. Elsevier, Amsterdam, pp. 15-31.

Crochemore, L., Perrin, C., Andréassian, V., Ehret, U., Seibert, S. P., Grimaldi, S., Gupta, H., Paturel, J. E. (2015). Comparing expert judgement and numerical criteria for hydrograph evaluation. Hydrological Sciences Journal, 60(3), 402423.

de Kok, J. L., Engelen, G., Maes, J. (2015). Reusability of model components for environmental simulation - Case studies for integrated coastal zone management. Environmental Modelling and Software, 68, 42–54.

Diez, E., McIntosh, B.S., 2009. A review of the factors which influence the use and usefulness of Information Systems. Environmental Modelling and Software 24 (5), 588-602.

Diez, E., & McIntosh, B. S. (2011). Organisational drivers for, constraints on and impacts of decision and information support tool use in desertification policy and management. Environmental modelling & software, 26(3), 317-327.

Dilling, L., Lemos, M.C. (2011). Creating Usable Science: Opportunities and Constraints for Climate Knowledge Use and Their Implications for Science Policy. Global Environmental Change, 21(2), 680–689.

Doherty, J., Hunt, R.J. (2009). Two statistics for evaluating parameter identifiability and error reduction. J. Hydrol. 366, 119–127. https://doi.org/10.1016/j.jhydrol.2008.12.018

Douthwaite, B., Kuby, T., van de Fliert, E., Schulz, S. (2003). Impact pathway evaluation: an approach for achieving and attributing impact in complex systems. Agricultural Systems, 78(2), 243-265. http://dx.doi.org/10.1016/S0308-521X(03)00128-8

Duggan, J. (2015). System Dynamics and Social_Ecological Systems Framework: Complimentary Methods for Exploring the Dynamics of Complex Systems. Systems Research and Behavioral Science, 32(4), 433-436.

Dunn, G., Laing, M., 2017. Policy-makers Perspectives on Credibility, Relevance and Legitimacy (CRELE). Environmental Science and Policy, 76, 146-152.

Elsawah, S., Guillaume, J.H.A., Filatova, T., Rook, J. and Jakeman, A.J. (2015) A methodology for eliciting, representing, and analysing stakeholder knowledge for decision making on complex socio-ecological systems: from cognitive maps to agent-based models. J Environmental Management, 151, 500-516.

Elsawah, S., McLucas, A., & Mazanov, J. (2017b). An empirical investigation into the learning effects of management flight simulators: A mental models approach. European Journal of Operational Research, 259(1), 262-272.

Elsawah, S., Pierce, S., Hamilton, S.H., van Delden, H., Haase, D., Elmahdi, A., Jakeman, A.J. (2017a) An overview of the System Dynamic process for integrated modelling of socio-ecological systems: Lessons on good modelling practice from five case studies. Environmental Modelling & Software, 93, 127145.

EPA. (2009). Guidance on the Development, Evaluation, and Application of Environmental Models. EPA/100/K-09/003. US Environmental Protection Agency, Washington DC. https://www.epa.gov/sites/production/files/201504/documents/cred_guidance_0309.pdf

Fienen, M.N., Doherty, J.E., Hunt, R.J., and Reeves, H.W., 2010, Using prediction uncertainty analysis to design hydrologic monitoring networks: Example applications from the Great Lakes water availability pilot project: U.S. Geological Survey Scientific Investigations Report 2010–5159, 44 p.

Folke, C., Hahn, T., Olsson, P., Norberg, J., (2005). Adaptive Governance of Social-Ecological Systems. Annual Review of Environmental Resources, 30, 441-473.

Forgie, V., Richardson, E., (2007). The community outcomes process and mediated modelling. International Journal of Sustainable Development, 10(4), pp.365-381.

Formetta, G., Antonello, A., Franceschi, S., David, O., Rigon, R., (2014). Hydrological modelling with components: A GIS-based open-source framework. Environmental Modelling & Software, 55, 190-200.

Franco, L. A., & Greiffenhagen, C. (2018). Making OR practice visible: Using ethnomethodology to analyse facilitated modelling workshops. European Journal of Operational Research, 265(2), 673-684.

Galelli, S., Humphrey, G. B., Maier, H. R., Castelletti, A., Dandy, G. C., & Gibbs, M. S. (2014). An evaluation framework for input variable selection algorithms for environmental data-driven models. Environmental Modelling & Software, 62, 33-51. doi:http://dx.doi.org/10.1016/j.envsoft.2014.08.015

Gilbert, G., Ahrweiler, P., Barbrook-Johnson, P., Narasimhan, K., & Wilkinson, H. (2018). Computational modelling of public policy: reflections on practice. Journal of Artificial Societies and Social Simulation, 21(1), 1-14.

Glynn, P. D., Voinov, A. A., Shapiro, C. D., White, P. A. (2017). From Data to Decisions: Processing Information, Biases, and Beliefs for Improved Manage of Natural Resources and Environments. Earth's Future, 5(4), 356-378.

Gneiting, T., Balabdaoui, F., Raftery, A.E., (2007). Probabilistic forecasts, calibration and sharpness. J. R. Stat. Soc. Ser. B (Statistical Methodol). 69, 243–268. https://doi.org/10.1111/j.1467-9868.2007.00587.x

Gober, P. A. (2018). Building Resilience for Uncertain Water Futures. London: Palgrave McMillian.

Goeller, B. F. (1988). A framework for evaluating success in systems analysis. In: Miser, H.J., Quade. E.S. (eds.) Handbook of system analysis: craft issues and procedural choices. John Wiley & Sons Ltd, p 567-618.

Gorddard, R., Colloff, M.J., Wise, R.M., Ware, D., Dunlop, M., (2016). Values, rules and knowledge: adaptation as change in the decision context. Environmental Science & Policy, 57, 60-69.

Gray, S., Paolisso, M., Jordan, R., Gray, S. (eds.) (2017) Environmental Modeling with Stakeholders: Theory, Methods, and Applications. Springer.

Guillaume, J.H.A., Hunt, R.J., Comunian, A., Blakers, R.S., Fu, B., (2016). Methods for Exploring Uncertainty in Groundwater Management Predictions, in: Jakeman, A.J., Barreteau, O., Hunt, R.J., Rinaudo, J.-D., Ross, A. (Eds.), Integrated Groundwater Management. Springer International Publishing, Cham, pp. 711–737. https://doi.org/10.1007/978-3-319-23576-9_28

Guillaume, J.H.A., Helgeson, C., Elsawah, S., Jakeman, A.J., Kummu, M., (2017). Toward best practice framing of uncertainty in scientific publications: A review of Water Resources Research abstracts. Water Resour. Res. https://doi.org/10.1002/2017WR020609

Haasnoot, M., van Deursen, W.P.A., Guillaume, J.H.A., Kwakkel, J.H., van Beek, E., Middelkoop, H. (2014). Fit for purpose? Building and evaluating a fast, integrated model for exploring water policy pathways. Environ. Model. Softw. 60, 99–120. https://doi.org/10.1016/j.envsoft.2014.05.020

Halbe, J., Pahl-Wostl, C., & Adamowski, J. (2018). A methodological framework to support the initiation, design and institutionalization of participatory modeling processes in water resources management. Journal of Hydrology, 556, 701716.

Hamilton, S.H., Guillaume, J., ElSawah, S., Jakeman, A.J. and Pierce, S.A. (2015) Integrated assessment and modelling: overview and synthesis of salient dimensions. Environmental Modelling and Software 64, 215-229.

Harmel, R.D., Smith, P.K., Migliaccio, K.W., Chaubey, I., Douglas-Mankin, K.R., Benham, B., Shukla, S., Muñoz-Carpena, R., Robson, B.J., 2014. Evaluating, interpreting, and communicating performance of hydrologic/water quality models considering intended use: A review and recommendations. Environmental Modelling & Software, 57, 40-51.

Hassenforder, E., Ducrot, R., Ferrand, N., Barreteau, O., Daniell, K. A., & Pittock, J. (2016a). Four challenges in selecting and implementing methods to monitor and evaluate participatory processes: Example from the Rwenzori region, Uganda. Journal of environmental management, 180, 504-516.

Hassenforder, E., Pittock, J., Barreteau, O., Daniell, K.A. and Ferrand, N. (2016b). The MEPPP framework: a framework for monitoring and evaluating participatory planning processes. Environmental management, 57(1), pp.7996.

Head, B. (2010). Wicked Problems in Water Governance: Paradigm Changes to Promote Water Sustainability and Address Planning Uncertainty." Urban Water Security Technical Alliance, Technical Report No. 38. http://www.urbanwateralliance.org.au/publications/UWSRA-tr38.pdf. Accessed May 29 2017.

Harvey, J. (Ed.). (1998). Evaluation cookbook. Edinburgh: Institute for Computer Based Learning.

Herman, J.D., Reed, P.M., Zeff, H.B., Characklis, G.W., (2015). How Should Robustness Be Defined for Water Systems Planning under Change? J. Water Resour. Plan. Manag. 04015012. https://doi.org/10.1061/(ASCE)WR.19435452.0000509

Holzworth, D.P., Huth, N.I., de Voil, P.G., (2010). Simplifying environmental model reuse. Environ. Model. Softw. 25, 269–275. https://doi.org/10.1016/j.envsoft.2008.10.018

Homès, B. (2011). Fundamentals of Software Testing, John Wiley & Sons, Incorporated.

Huitema, D., Meijerink, S., 2010. Realizing water transitions: the role of policy entrepreneurs in water policy change. Ecol. Soc. 15.

Hunka, A.D., Meli, M., Thit, A., Palmqvist, A., Thorbek, P. and Forbes, V.E., 2013. Stakeholders' perspective on ecological modeling in environmental risk assessment of pesticides: challenges and opportunities. Risk Analysis: An International Journal, 33(1), 68-79

Hutton, C., Wagener, T., Freer, J., Han, D., Duffy, C., & Arheimer, B. (2016). Most computational hydrology is not reproducible, so is it really science? Water Resources Research, 52(10), 7548–7555.

Ika, L.A., (2009). Project Success as a Topic in Project Management Journals. Proj. Manag. J. 40, 6–19. https://doi.org/10.1002/pmj.20137

Jahangirian, M., Taylor, S. J., Young, T., & Robinson, S. (2017). Key performance indicators for successful simulation projects. Journal of the Operational Research Society, 68(7), 747-765.

Jakeman, A.J., Letcher, R.A., Norton, J.P., 2006. Ten iterative steps in development and evaluation of environmental models. Environmental Modelling and Software 21, 602-614.

Jones, Natalie A., Pascal Perez, Thomas G. Measham, Gail J. Kelly, Patrick d'Aquino, Katherine A. Daniell, Anne Dray, and Nils Ferrand. "Evaluating participatory modeling: developing a framework for cross-case analysis." Environmental Management 44, no. 6 (2009): 1180.

Keen, P.G., 1980. Decision support systems: Translating analytic techniques into useful tools. Sloan Management Review, 21(3), 33.

Kelly, R.A., Jakeman, A.J., Barreteau, O., Borsuk, M.E., ElSawah, S., Hamilton, S.H., Henriksen, H.J., Kuikka, S., Maier, H.R., Rizzoli, A.E., van Delden, H. and Voinov, A.A. (2013). Selecting among five common modelling approaches for integrated environmental assessment and management. Environmental Modelling and Software 47, 159-181.

Klemeš, V. (1986). Operational testing of hydrological simulation models. Hydrological Sciences Journal, 31(1), 13-24.

Kolkman, D. A., Campo, P., Balke-Visser, T., & Gilbert, N. (2016). How to build models for government: criteria driving model acceptance in policymaking. Policy Sciences, 49(4), 489-504.
Kollmuss, A., Agyeman, J., (2002). Mind the Gap: Why do people act environmentally and what are the barriers to pro-environmental behavior?, Environmental Education Research, 8(3), 239-260.
Krause, P. and Flügel, W.A., 2005. Model integration and development of modular modelling systems. Advances in Geosciences, 4,1-2.
Krueger, T., Page, T., Hubacek, K., Smith, L., Hiscock, K. (2012). The role of expert opinion in environmental modelling. Environmental Modelling & Software, 36, 4-18.
Kunseler, E.-M., Tuinstra, W., Vasileiadou, E., Petersen, A. C. (2015). The reflective futures practitioner: balancing salience, credibility and legitimacy in generating foresight knowledge with stakeholders. Futures, 66, 1-12.
Laio, F., Tamea, S., (2006). Verification tools for probabilistic forecasts of continuous hydrological variables. Hydrol. Earth Syst. Sci. Discuss. 3, 2145– 2173. https://doi.org/10.5194/hessd-3-2145-2006
Larman, C., Basili, V. R. (2003). Iterative and Incremental Development: A Brief History. Computer 36(6), 47-56. http://dx.doi.org/10.1109/MC.2003.1204375
Larsen, K., Gunnarsson-Östling, U., (2009) Climate change scenarios and citizen-participation: Mitigation and adaptation perspectives in constructing sustainable futures. Habitat International 33(3), 260-266.
Lahtinen, T.J., Guillaume, J.H.A., Hämäläinen, R.P., (2017). Why pay attention to paths in the practice of environmental modelling? Environ. Model. Softw. 92, 74–81. https://doi.org/10.1016/j.envsoft.2017.02.019
Lockwood, M., Davidson, J., Curtis, A., Stratford, E., Griffth, R. (2010) Governance principles for natural resource management. Society and Natural Resources 23, 1-16.
Loucks, D.P., Van Beek, E., Stedinger, J. R. Dijkman, P., Jozef P.M., Villars, Monique, T. (2005). Water Resources Systems Planning and Management: An Introduction to Methods, Models, and Applications, UNESCO.
Lusiana, B., van Noordwijk, M., Suyamto, D., Mulia, R., Joshi, L., Cadisch, G. (2011) Users' perspectives on validity of a simulation model for natural resource management. International Journal of Agricultural Sustainability 9(2), 364-378
Maier, H.R., Guillaume, J.H.A., van Delden, H., Riddell, G.A., Haasnoot, M., Kwakkel, J.H., (2016). An uncertain future, deep uncertainty, scenarios, robustness and adaptation: How do they fit together? Environ. Model. Softw. 81, 154–164.
Matthews, K., Rivington, M., Blackstock, K., McCrum, G., Buchan, K., & Miller, D. G. (2011). Raising the bar?–The challenges of evaluating the outcomes of environmental modelling and software. Environmental Modelling & Software, 26(3), 247-257.
Martinez-Moyano, I. J., & Richardson, G. P. (2013). Best practices in system dynamics modeling. System Dynamics Review, 29(2), 102-123. doi:10.1002/sdr.1495
Matott, L. S., Babendreier, J. E., & Purucker, S. T. (2009). Evaluating uncertainty in integrated environmental models: A review of concepts and tools. Water Resources Research, 45(6), W06421. doi:10.1029/2008WR007301 Matthews, K., Rivington, M., Blackstock, K., McCrum, G., Buchan, K., & Miller, D. (2011). Raising the bar?–The challenges of evaluating the outcomes of environmental modelling and software. Environmental Modelling & Software, 26(3), 247-257.
Mayer, R.C., Davis, J.H., Schoorman, F.D. (1995). An integrative model of organisational trust. Academy of Management Review 20 (3), 709-734.
McConnell, S. (2004) Code Complete, 2nd edition. Microsoft Press, Washington, USA.
McIntosh, B.S., Ascough II, J.C., Twery, M., Chew, J., Elmahdi, A., Haase, D., Harou, J.J., Hepting, D., Cuddy, S., Jakeman, A.J., Chen, S., Kassahun, A., Lautenbach, S., Matthews, K., et al. (2011). Environmental decision support systems (EDSS) development - Challenges and best practices. Environmental Modelling and Software 26 (12), 1389-1402.
Mendoza, G., Cardwell, H. Guerrero, P. (2013) Integrated water resources management in Peru through shared vision planning. In: Griffiths and Lambert (Eds.) Free flow: reaching water security through cooperation. UNESCO/Tudor Rose, London, UK. pp 136–140.

Merritt, W.S., Fu, B., Ticehurst, J.L., El Sawah, S., Vigiak, O., Roberts, A.M,, Dyer, F., Pollino, C.A., Guillaume, J.H.A., Croke, B.F.W., Jakeman, A.J. (2017). Realizing modelling outcomes: A synthesis of success factors and their use in a retrospective analysis of 15 Australian water resource projects. Environmental Modelling & Software, 94, 63-72.

Mulligan, K.B. and Ahlfeld, D.P., (2016). Model reduction for combined surface water/groundwater management formulations. Environmental Modelling & Software, 81,102-110.

NRC (National Research Council) (2007). Evaluating Progress of the U.S. Climate Change Science Program: Methods and Preliminary Results. Washington, DC: The National Academies Press. https://www.nap.edu/download/11934. Accessed September 23 2017.

NRC (National Research Council) (2012) Assessing the Reliability of Complex Models: Mathematical and Statistical Foundations of Verification, Validation, and Uncertainty Quantification. Washington, DC: The National Academies Press. https://doi.org/10.17226/13395

Olsson, J.A., Andersson, L., (2006). Possibilities and problems with the use of models as a communication tool in water resource management. In Integrated Assessment of Water Resources and Global Change (pp. 97-110). Springer, Dordrecht.

Otaduy, I., Diaz, O., (2017). User acceptance testing for Agile-developed webbased applications: Empowering customers through wikis and mind maps. J. Syst. Softw. 133, 212–229. https://doi.org/10.1016/j.jss.2017.01.002

Pahl-Wostl, C. (2009). A Conceptual Framework for Analysing Adaptive Capacity and Multi-level Learning Processes in Resource Governance Regimes. Global Environmental Change, 19(3), 354-365.

Pahl-Wostl, C., Holtz, G., Kastens, B., Knieper, C., (2010). Analyzing Complex Water Governance Regimes: The Management and Transition Framework. Environmental Science & Policy, 13(7), 571-581.

Parker, W.S., (2009). II - Confirmation and adequacy-for-purpose in climate modelling. Proc. Aristot. Soc. Suppl. Vol. 83, 233–249. https://doi.org/10.1111/j.1467-8349.2009.00180.x

Patton, M. Q. (1996). A world larger than formative and summative. American Journal of Evaluation, 17, 131-144.

Perez, P., Aubert, S., Daré, W.S., Ducrot, R., Jones, N., Queste, J., Trébuil, G., Van Paassen, A., (2014) Assessment and monitoring of the effects of the ComMod approach. In: Companion Modelling. Springer, Dordrecht, pp. 155187.

Pianosi, F., Beven, K., Freer, J., Hall, J.W., Rougier, J., Stephenson, D.B., Wagener, T., (2016). Sensitivity analysis of environmental models: A systematic review with practical workflow. Environmental Modelling & Software, 79, 214-232.

Radhakrishnan, M., Islam, T., Ashley, R.M., Pathirana, A., Quan, N.H., Gersonius, B. and Zevenbergen, C., (2018). Context specific adaptation grammars for climate adaptation in urban areas. Environmental Modelling & Software, 102, 73-83.

Refsgaard, J.C., Henriksen, H.J., (2004). Modelling guidelines—terminology and guiding principles. Adv. Water Resour. 27, 71–82.

Refsgaard, J. C., van der Sluijs, J. P., Højberg, A. L., Vanrolleghem, P. A. (2007). Uncertainty in the environmental modelling process – A framework and guidance. Environmental Modelling & Software, 22(11), 1543-1556. Rittel H.W.J. and Webber M.M., Dilemmas in a general theory of planning. Policy Sciences 4(2), 1973, 155-169.

Robert, M., Thomas, A., Sekhar, M., Raynal, H., Casellas, É., Casel, P., Chabrier, P., Joannon, A. and Bergez, J.É., (2018). A dynamic model for water management at the farm level integrating strategic, tactical and operational decisions. *Environmental Modelling & Software*, *100*, 123-135.

Röckmann, C., Ulrich, C., Dreyer, M., Bell, E., Borodzicz, E., Haapasaari, P., Hauge, K.H., Howell, D., Mäntyniemi, S., Miller, D., Tserpes, G., Pastoors, M. (2012). The added value of participatory modelling in fisheries management – what has been learnt? Marine Policy, 36(5), 1072-1085.

Roughley, A. M. (2009). Developing and using program logic in natural resource management: user guide. Australian Government.

Rouwette, E. A., Vennix, J. A., van Mullekom, T. (2002). Group model building effectiveness: a review of assessment studies. System Dynamics Review, 18(1), 5-45.

Schuetz, T., Förch, W., Thornton, P., Vasileiou, I. (2017). Pathway to Impact: Supporting and Evaluating Enabling Environments for Research for Development. In: J. I. Uitto, J. Puri, & R. D. van den Berg (Eds.), Evaluating Climate Change Action for Sustainable Development (pp. 53-79). Cham: Springer International Publishing.

Schwanitz, V.J. (2013) Evaluating integrated assessment models of global climate change. Environmental Modelling & Software 50, 120-131.

Scriven, M. (1991). Beyond formative and summative evaluation. In: G. W. McLaughlin & D. C. Phillips (Eds.), Evaluation and education: At quarter century (pp. 19-64). Chicago, IL: University of Chicago Press.

Seidl, R. (2015). A functional-dynamic reflection on participatory processes in modeling projects. Ambio, 44(8), 750-765.

Sharp, E. and Curtis, A. (2014) Can NRM agencies rely on capable and effective staff to build trust in the agency? Australasian Journal of Environmental Management. 21: 3, 268-280.

Sharvelle, S., Dozier, A., Arabi, M. and Reichel, B., (2017). A geospatiallyenabled web tool for urban water demand forecasting and assessment of alternative urban water management strategies. *Environmental Modelling & Software*, *97*, 213-228.

Smajgl, A., & Ward, J. (2013). A framework to bridge science and policy in complex decision making arenas. Futures, 52, 52-58.

Smajgl, A., & Ward, J. (2015). Evaluating participatory research: framework, methods and implementation results. Journal of environmental management, 157, 311-319.

Smith, A.V., Sheppard, S.R.J., & Pinkerton, E.W. (2018) Community Forestry Practice and Visible Stewardship: A Case Study Evaluation in British Columbia. In: Gobster, Paul H.; Smardon, Richard C., eds. Visual resource stewardship conference proceedings: landscape and seascape management in a time of change. Gen. Tech. Rep. NRS-P-183. Newtown Square, PA: US Department of Agriculture, Forest Service, Northern Research Station, pp. 161-175.

Snow, A.D., Christensen, S.D., Swain, N.R., Nelson, E.J., Ames, D.P., Jones, N.L., Ding, D., Noman, N.S., David, C.H., Pappenberger, F., Zsoter, E., (2016). A High‐Resolution National‐Scale Hydrologic Forecast System from a Global Ensemble Land Surface Model. Journal of the American Water Resources Association, 52(4), 950-964.

Stave, K. (2010). Participatory system dynamics modeling for sustainable environmental management: Observations from four cases. Sustainability, 2(9), 2762-2784.

Stave, K. A., Beck, A., & Galvan, C. (2015). Improving learners' understanding of environmental accumulations through simulation. Simulation & Gaming, 46(3-4), 270-292.

Stein, B. A., Staudt, A., Cross, M. S., Dubois, N. S., Enquist, C., Griffis, R., Hansen, L. J., Hellmann, J. J., Lawler, J. J., Nelson, E. J., Pairis, A. (2013) Preparing for and Managing Change: Climate Adaptation for Biodiversity and Ecosystems. Frontiers of Ecological Environment, 11(9), 502-510.

Sterk, B., van Ittersum, M. K., & Leeuwis, C. (2011). How, when, and for what reasons does land use modelling contribute to societal problem solving? Environmental Modelling & Software 26(3), 310-316.

Stirling, A., 2010. Keep it complex. Nature 468, 1029–31. https://doi.org/10.1038/4681029a

Syme, G.J., Nancarrow, B.E., McCreddin, J.A. (1999). Defining the components of fairness in the allocation of water to environmental and human uses. Journal of Environmental Management 57, 51–70.

Tavella, E., & Franco, L. A. (2015). Dynamics of group knowledge production in facilitated modelling workshops: an exploratory study. Group Decision and Negotiation, 24(3), 451-475.

Syme, G. J., Sadler, B. S. (1994). Evaluation of Public Involvement in Water

Resources Planning: A Researcher-Practitioner Dialogue. Evaluation Review, 18(5), 523–542.

Ticehurst, J.L., Curtis, A., Merritt, W.S. (2011) Using Bayesian Networks to complement conventional analyses to explore landholder management of native vegetation. Environmental Modelling & Software, 26, 52-65.

Van Der Sluijs, J.P., Craye, M., Funtowicz, S., Kloprogge, P., Ravetz, J., Risbey, J., (2005). Combining Quantitative and Qualitative Measures of Uncertainty in Model-Based Environmental Assessment: The NUSAP System. Risk Anal. 25, 481–492.

van Voorn, G., Verburg, R., Kunseler, E.-M., Vader, J., Janssen, P. (2016). A checklist for model credibility, salience, and legitimacy to improve information transfer in environmental policy assessments. Environmental Modelling & Software, 83, 224-236.
Voinov, A., Bousquet, F. (2010). Modelling with stakeholders. Environmental Modelling & Software, 25(11), 1267-1488.
Voinov, A., Jenni, K., Gray, S., Kolagani, N., Glynn, P. D., Bommel, P., ... & Sterling, E. (2018). Tools and methods in participatory modeling: Selecting the right tool for the job. Environmental Modelling & Software, 109, 232-255.
Von Korff, Y., Daniell, K.A., Moellenkamp, S., Bots, P. and Bijlsma, R.M., 2012. Implementing participatory water management: recent advances in theory, practice, and evaluation. Ecology and Society, 17(1).
Ward, V., House, A., & Hamer, S. (2009). Developing a framework for transferring knowledge into action: a thematic analysis of the literature. Journal of Health Services Research & Policy, 14(3), 156-164.
Warren, K. (2014). Agile SD: Fast, Effective, Reliable. Paper presented at the Conference of the System Dynamics Society, Delft, Netherlands.
Weiss, C. H. (1995). Nothing as practical as good theory: exploring theorybased evaluation for comprehensive community initiatives for children and families. In: Connell et al. (ed) New Approaches to Evaluating Community Initiatives: Concepts, Methods, and Contexts. Washington, DC: Aspen Institute.
Westerveld, E. (2003). The Project Excellence Model: linking success criteria and critical success factors. International Journal of Project Management, 21, 411-418.
White, D. D., Corley, E. A., White, M. S. (2008) Water Managers' Perceptions of the Science–Policy Interface in Phoenix, Arizona: Implications for an Emerging Boundary Organization. Society and Natural Resources, 21, 230243.
White, D. D., Wutich, A., Larson, K. L., Gober, P., Lant, T., Senneville, C. (2010). Credibility, salience, and legitimacy of boundary objects: water managers' assessment of a simulation model in an immersive decision theater. Science and Public Policy, 37(3), 219.
William, D., Black, P. (1996). Meanings and consequences: a basis for distinguishing formative and summative functions of assessment? British Educational Research Journal, 22, 537-548.
Xiong, L., Wan, M., Wei, X., O'Connor, K.M. (2009). Indices for assessing the prediction bounds of hydrological models and application by generalised likelihood uncertainty estimation. Hydrol. Sci. J. 54, 852–871. https://doi.org/10.1623/hysj.54.5.852