

The evaluation of climate policy: theory and emerging practice in Europe

Dave Huitema · Andrew Jordan · Eric Massey · Tim Rayner · Harro van Asselt · Constanze Haug · Roger Hildingsson · Suvi Monni · Johannes Stripple

© The Author(s) 2011. This article is published with open access at Springerlink.com

Abstract Climate policy is a relatively young and dynamic area of public policy making. However, its development has attracted far more attention than the results it delivers in practice, which of course are the concern of policy evaluators. This article attempts to provide the first systematic cataloging of the emerging patterns of policy evaluation undertaken in different parts of the European Union. Theories of policy evaluation suggest that these evaluation practices should acknowledge the inherent complexity of climate policy making, be reflexive by questioning official policy goals, and be participatory. A meta-analysis of 259 climate policy evaluations suggests that current practice engages with some but not all of these issues. This article concludes by analyzing the implications of this finding for those in the academic and practitioner community who are keen to understand the extent to which climate policy evaluation is delivering on its promises.

Keywords Policy evaluation · Evaluation theory · Climate policy · European Union · Meta-analysis

Introduction

Climate change has been labeled a ‘wicked problem *par excellence*’ (Jordan et al. 2010), on account of its scientific complexity and the difficulty of securing agreement on policy

D. Huitema (✉) · E. Massey · H. van Asselt · C. Haug
Institute for Environmental Studies, VU University, De Boelelaan 1087, 1081 HV Amsterdam,
The Netherlands
e-mail: dave.huitema@ivm.vu.nl

A. Jordan · T. Rayner
Tyndall Centre for Climate Change Research, University of East Anglia, Norwich NR4 7TJ, UK

R. Hildingsson · J. Stripple
Department of Political Science, Lund University, Box 52, 22100 Lund, Sweden

S. Monni
Joint Research Centre - IES, European Union, Via E. Fermi 2749, 21027 Ispra, Italy

responses. The long timescales over which climate change manifests itself imply that the outcomes of actions taken now to mitigate its worst effects will not be observed for a considerable time. These and many other well-known difficulties have not, however, completely dissuaded national governments and international bodies from acting. Indeed, many have developed numerous climate policies, which consist of those policies aimed at reducing the impacts of anthropogenic climate change ('mitigation') and those seeking to reduce vulnerabilities or losses and capitalize on opportunities ('adaptation'). In Europe, the pace of national policy development quickened spectacularly after c. 1996 (a sixfold increase in policy activity; see Haug et al. 2008; Huitema et al. 2008). Meanwhile at European Union (EU) level, ambitious new targets have been set and many new policy instruments put in place (Jordan et al. 2010, forthcoming). These policy making activities, which link international, regional and national levels of governance, have attracted a fair amount of academic analysis (see for instance Gupta and Grubb 2000; Harris 2007; Schreurs and Tiberghien 2007) as have those in jurisdictions such as the USA and Australia, in which the pace of new policy development has been slower (Bailey and Marsh 2009; Paterson 2009).

Policy making is, of course, an important dimension of climate policy. But what of climate policy evaluation? Data collected for six EU states and for the EU as a whole reveal an eightfold increase in the number of evaluation reports produced in the period 2000–2005 (see Huitema et al. 2008). Other studies indicate that both academics and practitioners of climate policy are becoming more interested in evaluating the performance of existing policies (Haug et al. 2010). However, the decisions that originally informed these evaluation practices (i.e., the manner in which they were framed and performed) are still not very well described or understood.

The purpose of this article is to offer some novel insights into the emerging practices of climate policy evaluation in the EU. The evaluation of environmental policy has developed more slowly than in other policy realms such as welfare and education (Knaap and Kim 1998; Mickwitz and Birnbaum 2009). Key concepts are still not fully agreed upon, and many important methodological challenges remain (Mickwitz 2003). The aims of this article are therefore correspondingly modest: to investigate the number of evaluations done, the identity of the evaluators, the criteria they have used, the extent to which official policy goals are questioned in their evaluations, and the extent to which they incorporate or facilitate societal dialogue about the means and ends of climate policy.

In order to shed more light on these matters, this paper draws on the findings of the first systematic meta-analysis of climate policy evaluations undertaken anywhere in the world (see Haug et al. 2008; Huitema et al. 2008). A meta-analysis is a widely accepted method initially developed in the medical sciences, which aims to identify dominant patterns in the results of multiple assessments—or, for us, evaluations—conducted in a particular policy field (e.g., Glass et al. 1981; Stufflebeam 2001; Vedung 2005). The evaluations in the meta-analysis were compiled from database and internet searches, supplemented by 'snowballing' (i.e., asking different producers and users of evaluations—policy makers, evaluators, and so on—for suggestions). All of them are effectively in the public domain. They date from the period January 1998 to March 2007¹ and address the climate policies produced by six EU Member States—the United Kingdom, Germany, Italy, Finland, Portugal, and Poland. These broadly reflect the social, political and economic diversity

¹ These dates were chosen because evaluation reports prior to 1998 were very few in number. As the database was closed in April 2007, the most recent evaluations included in the analysis date from March 2007.

which is apparent in Europe, which in turn are commonly associated with different patterns of policy making and, one might assume, evaluation practices (see Furubo et al. 2002).² Finally, given the obvious importance of EU-level action, evaluations are included that examine the operation of policies covering the EU as a whole rather than just these six states.

Each evaluation was categorized along a series of 10 main criteria and 50 sub-criteria. The main criteria ranged from the affiliation of the author(s) and main sector(s) addressed, to whether the evaluation was technical-analytical (i.e., accepting official policy goals as given) or what will be termed reflexive (i.e., be prepared to question and continuously update official policy goals). From this long list, those studies that offered a systematic assessment of policies already in place (ex post evaluations) were identified and retained; those that were either not sufficiently systematic or that were wholly ex ante were excluded from the database.³ Policies were classified as relating to 'climate change' if they were reported as such in respective National Communications to the Secretariat of the United Nations Framework Convention on Climate Change (UNFCCC).⁴ This process resulted in a database of 259 evaluations. Before proceeding, it is important to be clear that they include both non-scientific evaluations (e.g., commissioned by NGOs and governments, etc.) and scientific evaluations (including books and peer-reviewed journal articles). In other words, the scope of the evaluation activity covered in the database is a relatively broad one, extending well beyond those commissioned by governments to evaluate the performance of their own policies or those of the EU.

Having established the broad aims and objectives of this article, the rest of the argument unfolds as follows. The next section introduces the broad topic of policy evaluation, notes the well-known distinction between rationalist and constructivist theoretical approaches, and argues that a middle ground between the two has begun to emerge. In the case of climate policy, in this middle ground, a number of critical issues are still being debated. Although that debate is far from settled, they imply that an evaluation should (1) be capable of acknowledging and handling the inherent complexity of climate policy making; (2) be reflexive in challenging both the means and the goals of a policy; and (3) be participatory in nature. The third section reports the findings of a meta-analysis of the 259 evaluations against these broad characteristics. Christie (2003) suggests there is an enormous gap between evaluation theory and evaluation practices in the USA. One of the purposes of this paper is to test whether this also holds for climate policy in Europe. The

² The two large northern Member States, the United Kingdom and Germany, are normally described as drivers of the European policy agenda (Jordan and Liefverink 2004), have robust economies, and display a relatively high degree of public involvement in policy making. The two southern European states, Portugal and Italy, while differing in size and economic conditions, have similar geographic characteristics, and both can be characterized as followers in the sphere of environmental policy. Finland, a small, rich and socially progressive country, represents the Scandinavian perspective, while Poland represents the block of post 2004 member states.

³ An evaluation was not considered to be systematic enough when it did not specify either which methods were used or what sources were drawn on. Using these criteria implied, a number of reports were excluded from the database, principally position papers from environmental NGOs.

⁴ All Parties to the UNFCCC are required to periodically submit National Communications to the UNFCCC Secretariat which contain information on emissions and removals of greenhouse gases and on the activities it has undertaken to implement the Convention. It is important to note therefore that the following analysis is based on information which has been *self-reported* by national governments (for more information, see Huitema et al. 2008 and http://unfccc.int/national_reports/items/1408.php).

final section concludes the analysis and points to some future research directions in this under-explored but politically very salient aspect of climate policy.

The main theories of policy evaluation

Evaluation: core meanings

In the mainstream literature, evaluation literally means ‘determination of value’ (Van de Graaf and Hoppe 1996; Vedung 2005), but there are many competing sub-definitions. One well-known definition states that policy evaluation is a ‘careful, retrospective assessment of merit, worth, and value of the administration, output and outcome of government interventions, which is intended to play a role in future practical action situations’ (Vedung 2005: 13). Another suggests evaluation is an ‘applied endeavour which uses multiple methods of inquiry and argument to produce and transform policy relevant information that may be utilized in political settings to resolve public problems’ (Dunn 2004: 35). Elsewhere, it has been defined as ‘a scientific evaluation of a certain policy area, the policies of which are assessed for certain criteria, and on the basis of which recommendations are formulated’ (Crabbé and Leroy 2008: 1).

These definitions differ in both obvious and more subtle ways. In our view, limiting the concept of policy evaluation to ‘scientific evaluation’ is not necessary as policy evaluation can be performed by non-scientists such as consultancy firms, lobby groups, and politicians. While policy evaluation is mainly geared toward public policy, it need not necessarily produce clear policy recommendations—the needs of other actors could also be targeted. Finally, one should be sensitive to the notion that the information produced by evaluation practices may or may not be utilized. However, demonstrating that a given evaluation has (or has not) had a direct and unambiguous policy impact is a notoriously difficult task (Weiss 1977; Nilsson et al. 2008).

Policy evaluation has gone through several stages since its emergence in the nineteenth century (Crabbé and Leroy 2008). Initially, its purpose was to assist national parliaments in assessing the lawfulness of government actions. After the Second World War, a period of immense state and policy development, the emphasis shifted to more administrative, managerial, and economic questions relating to the functioning of governments. And from the 1990s onwards, political questions related to public support for policies were increasingly asked. Different evaluation criteria emerged and became popular in these periods, but none of them has become redundant. Consequently today, would-be evaluators are confronted with a plethora to choose from. In the evaluation literature, the most mentioned are effectiveness and goal attainment, efficiency, cost-effectiveness, legitimacy, fairness, legal acceptability, and coordination with other policies (compare Dunn 2004; Vedung 2005; Wollmann 2007; Crabbé and Leroy 2008; Kraft and Furlong 2010). Table 1 summarizes our own understanding of these criteria. Later on, an attempt is made to explore which of them are actually being used in contemporary climate policy evaluation practices.

Theories of evaluation

The suggestion has been made that ‘evaluation as a field has not systematically developed or tested theory’ (King 2003: 57). The key word here is ‘systematically’ because there is certainly no lack of theory. Christie (2003), for instance, highlights no less than nine major

Table 1 Commonly used policy evaluation criteria

Criterion	Leading questions, examples
Goal attainment and effectiveness	Whether policy goals have been achieved and whether this can be attributed to the policy
Cost-effectiveness	How much of a given benefit is delivered per unit of expenditure, expressed as the net benefit or cost per unit of effectiveness? (e.g., tons of carbon mitigated or number of vulnerable people protected)
Efficiency	Have the right goals been formulated, should certain emission reductions should be achieved by one sector or another, or do the benefits of reduced emissions outweigh costs incurred?
Fairness	Relates to issues of equity, including the question whether ‘windfall profits’ (unfair competitive advantages) have arisen because of climate policies (e.g., emissions trading creates a profit potential for those with many emission credits, i.e., the bigger polluters)
Legitimacy	Does the public accept the policies, does the policy meet criteria of democratic accountability such as transparency?
Coordination	Is the policy well coordinated with existing other policies?
Legal acceptability	Are policies in accordance with legal principles?

approaches that build on starkly different assumptions. An important distinction is often made between more rationalistic and more constructivist approaches (see for example Sanderson 2002; Owens et al. 2004; Mickwitz and Birnbaum 2009). The rationalistic approach views policy as a means to achieve certain predefined goals. Policy development is mainly perceived as a task for a central actor, typically a state. Policies are therefore evaluated in order to inform new policy making practices. It is instrumental in the sense that it allows a principal to assess whether policy goals are being met or not (Abma and In ‘t Veld 2001; but see Pielke 2004). Collecting objective facts and describing the functioning of programs in light of democratically established goals are seen as vital tasks. This theoretical approach to policy evaluation has, however, been widely criticized for its limited usability, its tendency to be too uncritical of pre-established goals, its neglect of possible negative side effects, and its lack of stakeholder participation (Abma and In ‘t Veld 2001) (for a spirited defense, see Vedung 2005).

In response, a constructivist approach to evaluation started to develop. It stresses the autonomous character of policy, meaning that it tends to follow its own course of development. The goal of policy evaluation is therefore less to support the state and more to offer insights into the discourses and frames that are used by various actors to make sense of the world around them, including the nature of problems and the performance of policies. Guba and Lincoln (1981, 1989) for instance argue for a progressive development in policy evaluation practices away from measurement, first to description and then judgment, and culminating in ‘negotiation’. Here, the evaluation is viewed as an interactive process in which all participants have an equal say. The evaluator’s role is to be a process facilitator not the sole evaluator (Abma and In ‘t Veld 2001). The claims, concerns, and issues identified by stakeholders are thus at the very heart of the evaluation, *not* the goals embodied in the policy. Moreover, the outcomes of the evaluation are not the sole responsibility of the evaluator but are supposed to be debated between the evaluator and relevant stakeholders. Although a constructivist approach is often associated with learning, the assumption that it is the only way to promote it has been challenged. Owens et al. (2004: 1949–1950), for instance, have suggested that in some cases deliberation may

simply ‘excavate and expose the structure of the deadlock’. Conversely, ‘even quite technical procedures [may] have, as an unintended effect, provided important apertures for deliberation and learning’ (Owens et al. 2004: 1950).

A theoretical middle ground?

There are signs that a ‘middle ground’ between these two theoretical approaches is beginning to emerge (compare Pawson 2006), in which issues of *complexity*, *reflexivity*, and *participation* are very much center stage. To take the question of complexity first, the argument made by constructivists that modern-day problems are normatively and scientifically complex is increasingly acknowledged by both sides (see for example Hirschmüller and Hoppe 2001; Sanderson 2002). There is also a small but influential body of literature that attempts to link policy learning processes (which are always a factor in evaluations) to levels of complexity and the use of evaluation methods. As for reflexivity, Owens et al. (2004) suggest that the choice of approach should depend on the object of evaluation and the objective. Specialists command numerous legitimacy in addressing certain kinds of questions, but when there significant uncertainty and (most importantly) diverging framings of the problem, so the argument grows for using more deliberative approaches that encourage reflexivity (ibid.). Finally, forms of deliberation—always pushed very strongly by constructivists—are now increasingly being advocated by more rationalist scholars. Christie (2003) for instance concludes that nine leading scholars of evaluation, who subscribe to very different theoretical approaches, all accept that the involvement of stakeholders should be part of an evaluation process.⁵

Fischer (1995) has gone the furthest in trying to develop a formal synthesis of the two main approaches, which also combines the three issues of complexity, reflexivity, and participation. It uses the nature of the overriding question as the main sorting mechanism. He suggests that policy evaluation may occur at four levels: technical verification; situational validation; system vindication; and rational social choice. These levels differ mainly in reflexivity. Technical verification entails the evaluation of a policy for the purpose of asserting its empirical effectiveness. The central question often asked at this level is ‘does the policy achieve its stated goals’? This type of argumentation can be characterized as problem solving and according to Fischer is valid for issues that are not very complex. At the second level, situational validation, the analysis should determine whether the criteria used to judge the policy are themselves valid, e.g., ‘are the defined policy goals an adequate solution to the problem’? Here problem formulation and goal formulation and reflection upon them are critical. At the two higher levels, policy deliberation with high levels of participation concerns the justification and acceptability of the very value system adopted to judge the policy, sometimes coming down to core convictions about the preferred social order or ‘way of life’. The third level is system vindication, or general political argumentation, where the main issues revolve around the compatibility of the policy in question with accepted political values and general aims. As yet a higher level (rational social choice or ideological argumentation), core ideological debate is in place, where a fundamental change in life and the adoption of radically different social ideals seem necessary to achieve certain goals.⁶

⁵ However, participatory methods can be implemented in a variety of ways, as indicated by the distinction between Practical Participatory Evaluation and Transformative Participatory Evaluation (Cousins and Whitmore 1998).

⁶ Our summary here is based on Huitema et al. (2002).

Table 2 Four levels of evaluation

Level	Questions Typical evaluative criteria	Reflexive?
Technical verification	Are the policy goals achieved? <i>Effectiveness, cost-effectiveness</i>	No
Situational validation	Are the goals an adequate solution to the problem? <i>Allocative efficiency, coordination</i>	Yes
System vindication	Is the policy compatible with political values and accepted societal aims? <i>Fairness, legal acceptability, transparency, legitimacy</i>	Yes
Rational social choice	Is a fundamental change of life and new social ideals necessary? <i>Liberalism, capitalism, Marxism, sustainability, planetary survival, etc.</i>	Yes

Based on Fischer (1995)

Table 2 illustrates Fischer's line of reasoning. It specifically does not seek to prescribe any particular order and/or steps in 'handling' the three issues, but in general terms, he does seem to suggest that rationalist methods are better suited to resolving questions of a lower order, and constructivist methods to questions of a higher order.

Clearly, the jury is still out on whether and indeed how the two approaches can be synthesized into a 'middle ground'. Rather than approach that debate from a purely theoretical perspective, this paper starts from the perspective of everyday evaluation practices in order to see how they correspond to three salient issues, notably complexity, reflexivity, and participation. So to what extent do the 259 climate policy evaluations acknowledge and seek to make sense of complexity? Secondly, to what extent do they exhibit reflexivity in their willingness to challenge prevailing policy goals? And thirdly, in what ways do they incorporate the views of different stakeholders?

Before turning to the evaluations themselves, it is worth pausing to reflect on whether these three issues are at all relevant to the challenge of climate change.

The evaluation of climate policy

Starting first with *complexity*, the climate literature increasingly recognizes that the 'social-ecological' systems (Berkes and Folke 1998) that climate governors seek to govern exhibit many 'wicked' traits such as non-reducibility, spontaneity, and variability (Dryzek 1987). This makes surprise, unpredictability, and the possibility of unexpected 'tipping points' (Lenton et al. 2008) omnipresent. Related to that, fostering reflexivity is seen as being particularly important (e.g., Herrick and Sarewitz 2000). Goals that have been set (e.g., 20% emission reduction of CO₂ by the year 2020) may easily be overtaken by newly emerging scientific or societal consensus.

In the literature on climate governance, reflexivity is regularly held up as something to aspire to (see for example Hisschemöller and Hoppe 2001; Pahl-Wostl 2009; Haug and Huitema 2009; Huitema et al. 2009). This literature certainly emphasizes the need for reflexive evaluation practices (see for example Russel et al. 2010), which means that goals should be openly questioned rather than taken as given. Interestingly, the point that 'shallower' levels of evaluation may eventually lead to deeper and enduring forms of learning is hardly acknowledged. Whether this is because of the peculiarities of the climate

problem or because of a time lag between the general evaluation literature and the climate governance literature is a moot point. For the sake of argument, an assumption is made that greater reflexivity in evaluation is desirable and that it has a greater potential for stimulating learning than non-reflexive evaluation.

Finally, the climate policy literature contains indications that climate policy evaluation should be *participatory*. Climate governance is, by necessity, based on incomplete and uncertain information. It has been suggested that in such circumstances a clear division of labor between science (including evaluators) and policy is near impossible as this would inhibit reflexive governance and limit the capacity of all stakeholders to deal with uncertainty and change (e.g., Moberg and Galaz 2005; Pielke 2007). Rather, they can—and indeed should—‘learn together to manage together’ (Ridder et al. 2005). Participatory evaluation is thus seen as being even more of a necessity than in other areas of environmental public policy (see for example Coenen et al. 1998; Mickwitz 2003). The stakeholders to be involved could be ordinary citizens, politicians and other decision makers, private companies and local implementing officials, etc. (see Vedung 2005: 71).

In short, the three issues do seem very relevant to climate change. Having discussed policy evaluation in general theoretical terms, the next section analyzes the emerging practices of climate policy evaluation.

Climate policy evaluation practices in Europe

How many evaluations are there?

The first thing to note is that evaluation practices vary greatly across the EU, with the UK taking the lead (78 evaluations of UK climate policies⁷); by contrast, a relatively low number of evaluations have been produced in Portugal and Poland (10 and 6 evaluations of their policies, respectively⁸). Evaluation activity at the EU level is quite high (105 evaluations of EU policies). In terms of timing, the number of evaluations until 2001 was fairly constant (around 10 p.a.) but then rose quickly to over 20 p.a. in 2002, before peaking at above 80 p.a. in 2006 (see Fig. 1).

Who has produced them?

Who produced these 259 evaluations? In principle, anyone is free to perform evaluations and publish the results. As shown by Fig. 2, universities and independent research institutes, followed by consultancy firms, are the most active evaluators. International and national governmental bodies are also relatively active. NGOs and industry groups appear to have performed relatively few evaluations, which may be due to lack of resources or will on the part of NGOs to produce and disseminate evaluations.

The majority of the evaluations (58%) were not commissioned, suggesting a certain level of independence on the part of the evaluators.⁹ A relatively small share (34%) was

⁷ This includes all evaluations of national policies, including those made by organizations based outside the country.

⁸ *Idem*.

⁹ The category of non-commissioned studies here includes EU-funded research projects, which are often instigated by DG Research and therefore operate largely independently from those responsible for policy development.

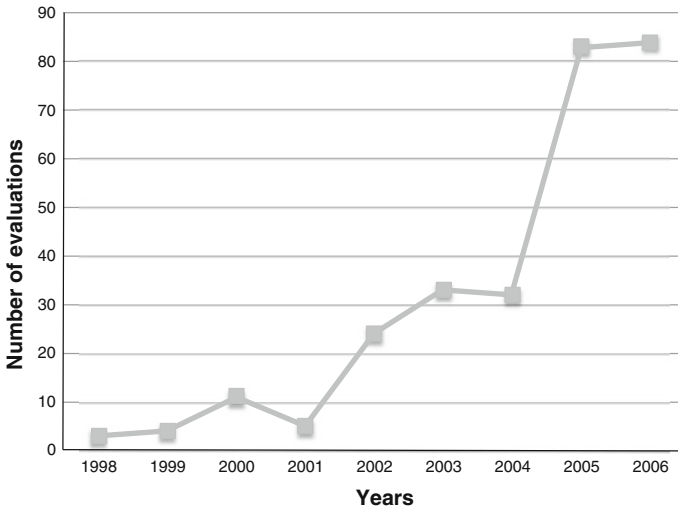


Fig. 1 The number of evaluations published annually (1998–2006) in the six EU Member States and the EU. The figure does not show data for 2007 as the analysis includes only reports that were produced until April 1, 2007

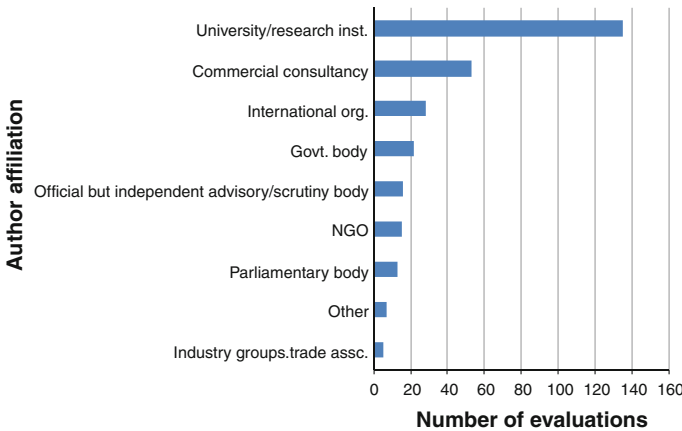


Fig. 2 Affiliation of authors per category. The total number is greater than the total number of evaluations due to the fact that many evaluations are authored jointly

commissioned. For a surprisingly large share (8%), it was unclear whether they had been commissioned or not. Of those that were commissioned, the most active commissioning agents were central governments (59% of the total), followed by international organizations, NGOs, and industry groups (12, 10, and 9%, respectively).

Which evaluation criteria have been used?

As Fig. 3 indicates, the three most commonly used criteria were effectiveness and/or goal achievement, efficiency, and cost-effectiveness (used in 213, 74, and 72 evaluations,

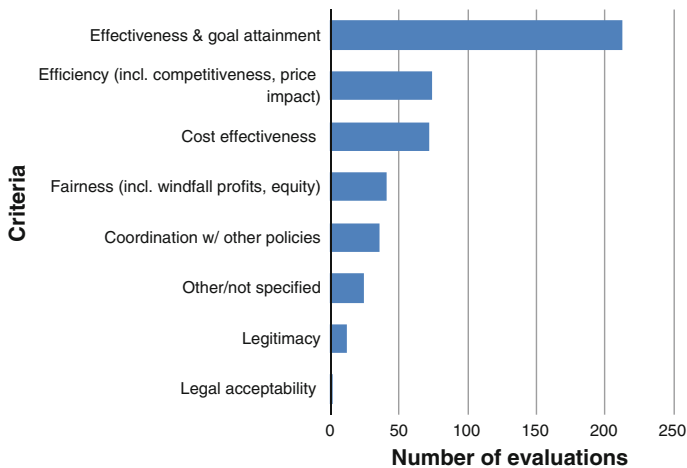


Fig. 3 Criteria used in the evaluation reports. Many evaluations use multiple criteria; hence, the total number used exceeds the number of evaluations

respectively). Other criteria such as fairness, coordination with other policies, and legitimacy were used far less frequently.

Is complexity acknowledged?

Awareness of complexity can be demonstrated in a number of ways, notably attention to unwanted side effects of policy interventions, the use of multiple criteria and methods, and the reconstruction of so-called intervention theories, i.e., the ideas about cause and effect which informed the policies (for a discussion of this, see Mickwitz 2003¹⁰). For the sake of simplicity, only data for the use of multiple methods and criteria were collected. Figures 4 and 5 show how diverse the set of methodologies and criteria used really is.

As for the number of methods used in the various evaluations, the data show that the use of multiple methodologies is not rare, but that more than half of the evaluations use only one methodology. As for the number of criteria per study, the average number is 1.8 (473 criteria over 259 evaluations). The largest group of evaluations uses only one criterion, whereas evaluations that employ more than three criteria are rare. There is no established scale to judge these findings, but it seems reasonable to conclude that the use of multiple methods and multiple criteria has not yet caught on in practice.

Reflexivity: do evaluations challenge established goals?

If policy evaluations questioned official policy goals, they were categorized as reflexive and vice versa. The overwhelming majority (82%) of the evaluations were non-reflexive in their outlook. Given the low threshold used (any sign that a policy was questioned was deemed sufficient) and the clear theoretical inclination toward reflexive evaluation noted above, one might have expected a rather higher percentage.

¹⁰ Mickwitz (2003), citing others, suggests different forms of triangulation to deal with the 'impact problem'. The term 'impact problem', as Mickwitz uses it, refers to the fact that there are typically long and complex chains linking policy interventions to changes in the environment. To assess and evaluate impacts, evaluators should not therefore rely on one single criterion or method.

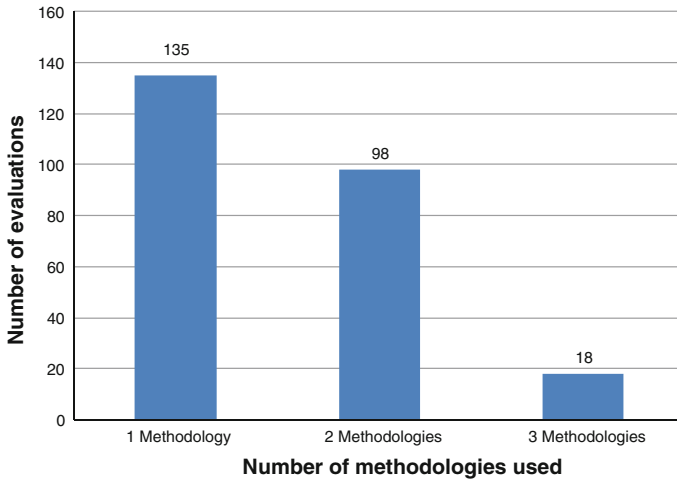


Fig. 4 Number of methodologies used in the evaluations

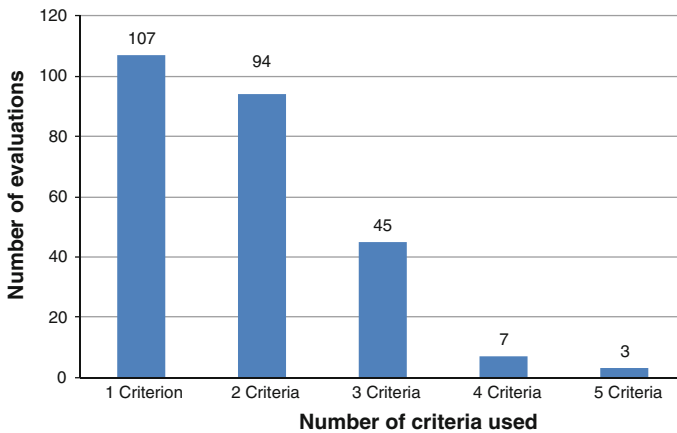


Fig. 5 Number of criteria used in evaluations. Some evaluations are lacking explicitly stated criteria

In an attempt to assess the degree of reflexivity among these 18%, the evaluation criteria used by the authors were analyzed. Criteria that are connected to Fischer's second level ('situational validation', see Table 2), efficiency and coordination, are used about twice as often (24 out of 46 reflexive reports) as those connected with his third level ('system vindication'), notably legal acceptability, legitimacy, and fairness (13 out of 46 reflexive reports). No evaluations were found that targeted the fourth level, that of social rational choice. Consequently, even among reflexive reports, the level of reflexivity observed was relatively low.

Clearly, a complex problem like climate change has to be simplified to render it amenable to policy making. Institutionally, the typical strategy is to disaggregate cross-cutting problems like climate change into a range of sectoral responses managed by individual line ministries (Jordan and Lenschow 2008). However, policy theory tells us that

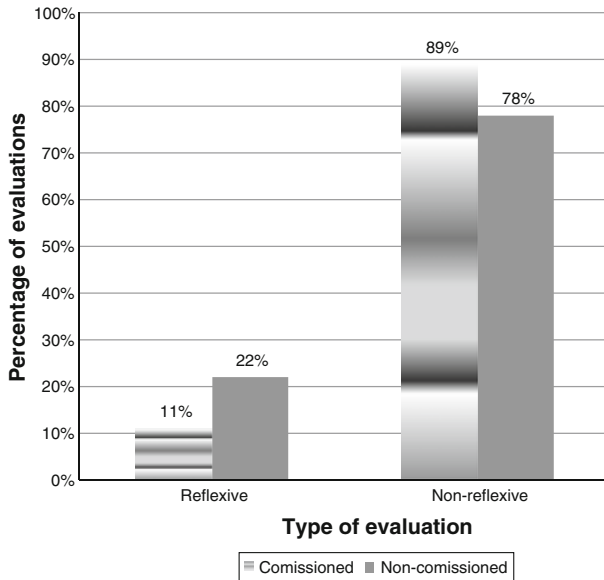


Fig. 6 Reflexive vs. technical reports for commissioned and non-commissioned evaluations

these ministries and their associated networks rarely challenge their core policy theories and belief systems, etc. (Huitema and Meijerink 2009). The meta-analysis of climate policy evaluation practices demonstrates how far this tendency to simplify has permeated the practices of evaluation, which in principle could be a valuable means to reveal and better appreciate (rather than obscure) the underlying complexity of policy problems.

This is not to suggest that evaluation has to look at everything. Key uncertainties have to be bracketed off because, where they not, there would be a real danger of ‘paralysis by analysis’. One way to assess how complexity was handled is to look at whether there is a difference between commissioned and non-commissioned reports and chiefly whether reflexivity is correlated with certain types of authors, or with certain methodologies. Figure 6 shows that the commissioned reports in the database were less reflexive in nature than non-commissioned reports. In fact, the share of reflexive reports is twice as high for non-commissioned reports as for commissioned reports.

One interesting hypothesis is that authorship is also correlated with reflexivity. This is possibly revealed in Fig. 7. Of note is the relatively large percentage of reflexive reports authored by parliamentary bodies—16% compared with their overall contribution to evaluation of 4%. The proportions for NGO-authored reports are very similar.

Finally, an attempt was made to examine how far reflexive and non-reflexive evaluations differed in terms of the methods employed. This reveals that, compared to the overall sample, reflexive reports are much less often based on modeling, and more often apply document analysis and participatory methods (see Fig. 8). The extent to which such evaluations employ documentary analysis is somewhat surprising given the theoretical claims made about reflexivity being enhanced through using participatory methods. This finding about climate policy evaluation practices is in line with the argument of authors who suggest that the relationship between reflexivity and participatory methods is more complex than often assumed (Owens et al. 2004; Lehtonen 2006).

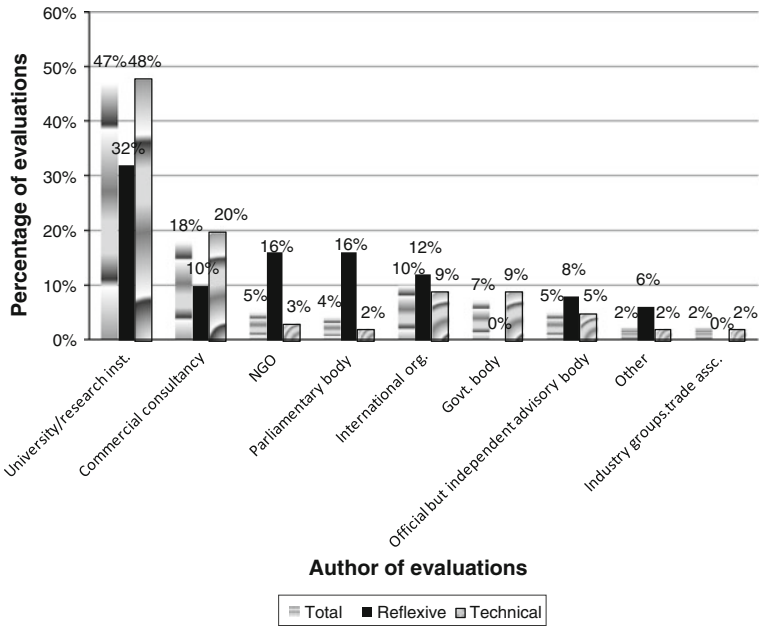


Fig. 7 Reflexivity and authorship

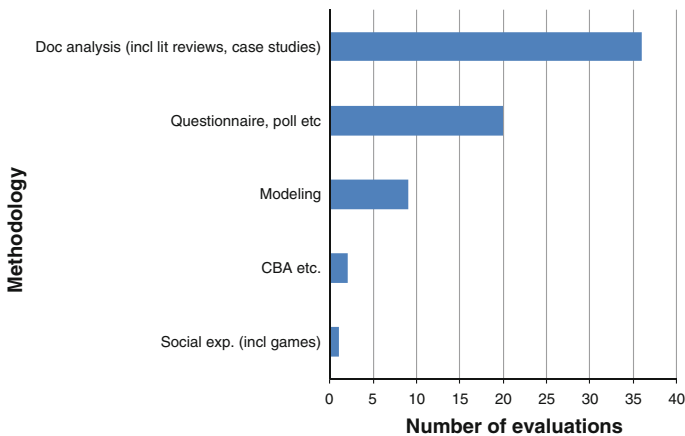


Fig. 8 Percentage of reflexive reports that employ each methodology. Most reports employ more than one methodology

How participatory are evaluations?

Turning finally to participation, this aspect was assessed in two ways. First, by looking at the methodologies applied in the various evaluations, in particular whether any sought to involve different stakeholders (NGOs, target groups, industry representatives, etc.) in the framing of the evaluation. Figure 9 shows that ‘document analysis’ is the most popular method by far; it was applied in about half of the 259 evaluations. This category includes

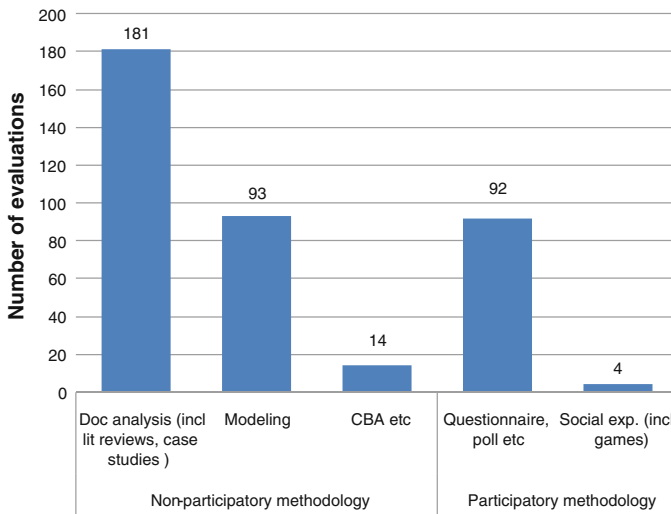


Fig. 9 Methodologies employed. The combined number of methodologies listed here is greater than the total number of evaluations because some evaluations applied multiple methodologies

literature reviews, analysis of legislation, and (secondary analysis of) case studies and monitoring data (involving no additional modeling). Two other popular methodologies are modeling and questionnaires; they were both applied in over a quarter of the evaluations in our sample.

For the sake of simplicity, three methodologies were categorized as being ‘participatory’: social experiments; stakeholder interviews; and questionnaires; two other methods, namely modeling¹¹ and cost–benefit analysis, were assumed to be ‘non-participatory’.¹² This assumption leads to the finding that 75% of all methods selected for use in our sample of 259 evaluations are non-participatory, and only 25% are participatory. While these findings are interesting, it should be recalled that these numbers include evaluations that employed multiple methodologies. Looking at the evaluations that employed only one methodology (i.e., the majority of the evaluations) only 8% (11/135) used participatory methods. For evaluations employing two methodologies, 67% (66/98) evaluations had participatory methods, and for evaluations employing three methodologies, 94% (17/18) used participatory methods. This finding means that participatory methods may not be the first choice of evaluators, but that they do use them as soon as the number of methodologies expands.

Secondly, an assessment was made of whether or not the evaluations indicated that representatives of industry, pressure group campaigners, government officials or citizens, etc. were directly involved, beyond simply acting as interviewees and/or survey respondents, etc. Using this criterion, only 12 studies (4.2%) could be said to be clearly participatory in nature. Remarkably, 11 of these were undertaken not at national, but at EU level. In the 12, a broad range of stakeholders were involved, notably industry (12/12),

¹¹ Modeling includes bottom-up simulation modeling of the energy sector, macroeconomic modeling, numerical simulations of market development, dynamic energy systems modeling.

¹² This is of course a slight over-simplification. We realize that most methodologies can be applied in more or less participatory ways. We are also aware of the fact that the degree of participation is limited in most of the methodologies we have labeled as ‘participatory’.

environmental NGOs (12/12), research organizations (8/12), government officials (9/12), and individuals (1/12). In most cases, these stakeholders were involved in describing the current situation, or in the concluding and policy recommendation sections. Crucially, they were not involved in determining the precise terms of reference of evaluations or specifying the evaluation criteria.

Discussion and conclusions

Climate policy is a young and dynamic area of policy making which began in the 1980s and then really took off in Europe in the 1990s. This pattern of development, combined with the underlying desire for more evidence based policy making, makes the topic of climate policy evaluation extremely important for both academics and policy makers. This article has drawn on a meta-analysis of 259 climate policy evaluations from six European countries and the EU to analyze how the practices of evaluation have evolved. The evidence collected suggests that evaluation is following a similar but slightly lagged pattern to that of policy making, with a pronounced growth after 2001. University researchers, independent research institutes, and consultancy firms are the most active evaluators. The large representation of university researchers is due to the fact that scientific articles were included in the database of evaluations. The majority of the evaluations examined (58%) were not commissioned, and only 34% were commissioned. For a surprisingly large share of studies (8%), it was not possible to determine whether they had been commissioned or not, which suggests that the transparency of evaluation practices is not as high as it could be. In terms of criteria, goal achievement and effectiveness were by far the most common criteria used (213/259 used one or both of these), followed at a considerable distance by efficiency and cost-effectiveness (which were used in 74 and 72, respectively).

In section “[The main theories of policy evaluation](#)”, rationalistic and a more constructivist approaches to evaluation were identified. These two approaches differ in terms of their view of policy; conception of science-policy relations; appropriate methodologies and their underlying ontologies and epistemologies. In theory, there are signs of an emerging middle ground between these two approaches. In this middle ground, one finds an active debate about how to handle complexity, reflexivity, and the use of participatory methods. There are signs of a convergence in the way evaluation is theorized; but are evaluation practices following the same path?

Evaluation theory tells us that *complexity* requires the application of multiple criteria and the use of various evaluation methods. However, the vast majority of current evaluations do not sufficiently acknowledge complexity. *Reflexivity* was defined in terms of the willingness to question formal policy goals and was further differentiated into different levels based on the numbers of criteria used in evaluation analyses. The majority of evaluations were found not to be reflexive. Moreover, those evaluations that are reflexive are only weakly reflexive in the sense that they remain at relatively low levels in Fischer’s typology. These findings are very salient because policy evaluation theory (and indeed related work on participatory governance) suggests that reflexive and participatory approaches are a very important means for society to ‘learn its way out of a problem’ like climate change. Finally, in terms of *participation*, an attempt was made to quantify the kinds of methodologies used and the extent to which the involvement of stakeholders went further than being an object of study. Measured in those ways, the overwhelming majority of the evaluations in the database do not meet the basic criterion of a participatory analysis (up to 95.8%, depending on method of measurement).

In summary, there is a sizeable gap between evaluation theory and evaluation practices—something that has been observed in other parts of the world (Christie 2003). Thus, there has been very little practical convergence toward the theoretical ‘middle ground’ noted above. It would be interesting to find out what has caused this gap between theory and practice. If similar dynamics are indeed present in Europe as in the USA (Christie 2003), then perhaps evaluators’ perceptions of themselves (for instance, do they think they know enough about stakeholders not to need participatory methods?) and/or their desire to be seen as objective (which could conceivably lead them to apply more quantitative techniques) might play a role.

According to Lehtonen (2005: 169) and Martinuzzi (2004), evaluation practices have the potential to act as a new form of environmental governance. The analysis presented here suggests that there is still a very long way to go before this potential is fully realized. Indeed, as the practices of climate policy evaluation continue to develop, further meta-analyses such as this could be a useful way to establish how far the situation has changed (i.e., facilitate reflexive thinking at a more meta-level).

We conclude by highlighting several limitations to our analysis and, partly in line with that, some potential avenues for future work. One clear limitation to our work is that we have only looked at six Member States, although at present there is nothing of comparable scope and depth. This means that evaluation practices in 21 Member States remain unexplored. Had they been included how would it have affected the outcomes? Depending on which countries are added, it could affect the findings quite significantly. Evidently, countries such as the UK and Germany are gradually developing more mature evaluation systems, whereas policy evaluation in countries such as Italy and Portugal largely depend on pressure from international organizations such as the EU (see Furubo et al. 2002).

A second set of limitations relates to the fact that we have only looked at the ‘supply side’ of evaluation; we have not really opened up the ‘demand side’, other than to note who is commissioning evaluations and broadly for what purpose. Other than this, we have not analyzed the commissioning process or analyzed how far the information and knowledge in the evaluations that have been commissioned have affected policy makers’ understandings and, ultimately, their behavior. The conceptual tools to do this are certainly available (see for example Weiss 1977). It would be interesting to determine whether emerging practices are feeding through to policy learning (Haug et al., forthcoming), through what routes and with what effects. It would also be useful to analyze the commissioning practices in more depth, given the rhetorical commitment to more evidence-based policy making. The findings (reported above) that commissioned work is generally less reflexive than non-commissioned work and that academics produce less reflexive work than NGOs and parliamentary bodies do certainly warrant further research.

We should also raise some issues in connection with the more theoretical matters discussed above. Complexity, reflexivity, and participation are key features of contemporary evaluation theory. But just how complex are climate problems, compared to other contemporary policy issues such as agriculture or development assistance? By calling all climate issues ‘complex’, we are in danger of possibly overlooking certain aspects that are less complex (for instance, the fact that emission reductions can often easily be achieved in combination with reduced expenditures by increasing energy efficiency) or might become so over time (for instance, because fewer countries continue to contest the existence and causes of climate change). The fact that we have so much policy activity in the climate domain possibly suggests that what was once a very unstructured situation is becoming more structured. But this leads to many other questions. There are various indications in this article (both in the theoretical and in the empirical parts) that suggest that complexity,

reflexivity, and participatory analysis have a more complicated relation than is often assumed and that this should be the subject of more work. For instance, our data suggest that the application of participatory methods does not automatically imply reflexivity. Therefore, we see a great need for conceptual and empirical work on complexity and its measurement, a need for analysis that better examines the implications of complexity for evaluation, and for work that better connects reflexivity and participation.

Nonetheless, our findings do inform a new research agenda focusing on what could be termed the ‘politics of environmental evaluation’. This agenda should embrace the highly differentiated practices of evaluation identified in this article. As indicated in our introduction, we were relatively broad in our definition of what constituted an evaluation, which meant that we also included journal articles and other academic publications. Some, especially those who associate evaluation with commissioned research by bureaucracies or parliaments, may argue that these do not count as ‘real’ evaluations. However, using such a broad selection did allow us to make some observations that would have otherwise been impossible. Specifically, we found that non-commissioned evaluations were twice as likely to contain a reflexive element as commissioned ones. If we combine this finding with the fact that governmental bodies are the most active commissioning agents, the explanation is likely to be that such bodies—which often have a specified policy agenda—are less keen on reflexive evaluations than other bodies. Equally interesting is the finding that university researchers are less likely to produce reflexive evaluations than what could be expected on the basis of their share in the total amount of evaluations and the fact that they (supposedly) enjoy an independent position. Instead, parliamentary committees, international organizations, and NGOs are much more supportive of reflexive evaluation.

The more comparative dimensions of our analysis are also worthy of further analysis, for example the relationship between political leadership and evaluation practices. For example, are the so-called environmental lead states that routinely push for higher environmental standards (Jordan and Lenschow 2008) also the more active evaluators? Are they as supportive of a greater role for the EU in these matters as they are in relation to policy making? Are the environmental lead states more supportive of greater levels of reflexivity? On the one hand they may, insofar as they are more interested in policy ‘improvement’. On the other hand, it may be quite difficult to criticize goals that are already relatively ambitious. Finally, has the EU’s involvement gradually harmonized evaluation practices (as it has done in relation to policy making) or do national differences in approach remain? Indeed, is there evidence that evaluation practices are gradually centralizing in EU-level bodies such as the European Environment Agency?

Derlien and Rist (2002) suggest that the EU is affecting the formation of national evaluation cultures in the Member States. One question one could ask is whether the EU’s preference for more participatory evaluations is having a trickledown effect on them. Derlien and Rist also argue that in the 1990s, the emphasis of evaluation in many countries has gradually shifted from the provision of information to more of an allocation function. By this, they mean that evaluation is increasingly used not only to improve policies (as it was in the 1960s and 1970s) but rather to decide whether or not to keep them. Is the increased level of evaluation activity detected in this paper really part of a wider effort to improve policy by learning about its effects (thus indicating a continuation of older evaluation traditions) or really an effort to keep them in check? Sadly, these very salient questions go well beyond the scope of this particular article, but they do suggest that the field of policy evaluation is absolutely ripe for new, empirically informed comparative work.

Acknowledgments The authors kindly acknowledge the support of the European Commission for the research reported here (ADAM project—contract GOCE-018476; the Livediverse project—FP7 contract 211392). The anonymous reviewers provided numerous constructive suggestions, and we are very grateful for the way in which they helped us improve this article.

Open Access This article is distributed under the terms of the Creative Commons Attribution Noncommercial License which permits any noncommercial use, distribution, and reproduction in any medium, provided the original author(s) and source are credited.

References

- Abma, T., & In 't Veld, R. (Eds.) (2001). *Handboek beleidswetenschap*. Amsterdam, The Netherlands: Boom Uitgevers.
- Bailey, I., & Marsh, S. (2009). Facing up to the greenhouse challenge? Australian climate politics. In H. Compston & I. Bailey (Eds.) *Turning down the heat* (pp. 202–222). Basingstoke, UK: Palgrave Macmillan.
- Berkes, F., & Folke, C. (Eds.) (1998). *Linking social and ecological systems: Management practices and social mechanisms for building resilience*. Cambridge, UK: Cambridge University Press.
- Christie, C. (2003). What guides evaluation? A study of how evaluation practice maps onto evaluation theory. *New Directions for Evaluation*, 97, 7–35.
- Coenen, F. H. J. M., Huitema, D., & O'Toole, L. J., Jr. (Eds.) (1998). *Participation and the quality of environmental decision making*. Dordrecht: Kluwer Academic.
- Cousins, J. B., & Whitmore, E. (1998). Framing participatory evaluation. *New Directions for Evaluation*, 80, 87–105.
- Crabbé, A., & Leroy, P. (2008). *The handbook of environmental policy evaluation*. London, UK: Earthscan.
- Derlien, H.-U., & Rist, R. C. (2002). Policy evaluation in international comparison. In J. E. Furubo, R. C. Rist, & R. Sandahl (Eds.) *International atlas of evaluation* (pp. 439–455). New Brunswick, NJ: Transaction Publishers.
- Dryzek, J. S. (1987). *Rational ecology: Environment and political economy*. New York, USA: Basil Blackwell.
- Dunn, W. (2004). *Public policy analysis* (3rd ed.) Englewood Cliffs, NJ: Prentice Hall.
- Fischer, F. (1995). *Evaluating public policy*. Chicago, USA: Nelson-Hall Publishers.
- Furubo, J. E., Rist, R. C., & Sandahl, R. (Eds.) (2002). *International atlas of evaluation*. New Brunswick, NJ: Transaction Publishers.
- Glass, G. V., McGaw, B., & Smith, M. L. (1981). *Meta-analysis in social research*. Beverly Hills, USA: Sage Publications.
- Guba, E., & Lincoln, Y. (1981). *Effective evaluation*. San Francisco, CA: Jossey-Bass.
- Guba, E., & Lincoln, Y. (1989). *Fourth generation evaluation*. London, UK: Sage.
- Gupta, J., & Grubb, M. (Eds.) (2000). *Climate change and European leadership: A sustainable role for Europe?*. Cheltenham, UK: Edward Elgar.
- Harris, P. G. (Ed.) (2007). *Europe and global climate change: Politics, foreign policy and regional cooperation*. Cheltenham, UK: Edward Elgar.
- Haug, C., & Huitema, D. (2009). Leren van een beleidsexercitie. De casus van het Europese milieubeleid. *Bestuurskunde*, 18(3), 36–46.
- Haug, C., Huitema, D., & Wenzler, I. (forthcoming). Learning through games? Evaluating the learning effect of a policy exercise on European climate policy. *Technological Forecasting & Social Change*, 78. doi:10.1016/j.techfore.2010.12.001.
- Haug, C., Rayner, T., Huitema, D., Massey, E., Monni, S., van Asselt, H., et al. (2008). *What makes climate policies effective? Findings from a 'meta analysis' of recent policy evaluations*. Institute for Environmental Studies (IVM). Available at <http://adamproject.info/index.php/Downloads/>.
- Haug, C., Rayner, T., Jordan, A., Hildingsson, R., Strippel, J., Monni, S., et al. (2010). Navigating the dilemmas of climate policy in Europe: evidence from policy evaluation studies. *Climatic Change*, 101(3–4), 427–445.
- Herrick, C., & Sarewitz, D. (2000). Ex post evaluation. A more effective role for scientific assessments in environmental policy. *Science, Technology and Human Values*, 25(3), 309–331.
- Hisschemöller, M., & Hoppe, R. (2001). Coping with intractable controversies: The case for problem structuring in policy design and analysis. In M. Hisschemöller, W. Dunn, R. Hoppe, & J. Ravetz (Eds.) *Knowledge, power and participation in environmental policy analysis*. *Policy Studies Review Annual*, 12, 47–72.

- Huitema, D., Jeliaskova, M., & Westerheijden, D. F. (2002). Phases, levels and circles in policy development: The cases of higher education and environmental quality assurance. *Higher Education Policy*, 15(2), 197–215.
- Huitema, D., & Meijerink, S. (Eds.) (2009). *Water policy entrepreneurs. A research companion to water transitions around the globe*. Cheltenham, UK: Edward Elgar.
- Huitema, D., Mostert, E., Egas, W., Moellenkamp, S., Pahl-Wostl, C., & Yalcin, R. (2009). Adaptive water governance. Assessing adaptive management from a governance perspective. *Ecology and Society*, 4(1), 26 [online]. <http://www.ecologyandsociety.org/vol14/iss1/art26/>.
- Huitema, D., Rayner, T., Massey, E., Haug, C., Hildingsson, R. Monni, S., et al. (2008). *Climate policy evaluation across Europe*. Institute for Environmental Studies (IVM). Available at <http://adamproject.info/index.php/Downloads/>.
- Jordan, A. J., Huitema, D., van Asselt, H., Rayner, T., & Berkhout, F. (Eds.) (2010). *Climate change policy in the European Union. Confronting the dilemmas of mitigation and adaptation?* Cambridge, UK: Cambridge University Press.
- Jordan, A. J., & Lenschow, A. (Eds.) (2008). *Innovation in environmental policy*. Cheltenham, UK: Edward Elgar.
- Jordan, A. J., & Liefferink, D. (Eds.) (2004). *Environmental policy in the European Union*. London, UK: Routledge.
- Jordan, A. J., Van Asselt, H., Berkhout, F., Huitema, D., & Rayner, T. (forthcoming). Climate change policy in the European Union: Understanding the paradoxes of multi-level governing. *Global Environmental Politics*, 11.
- King, J. A. (2003). The challenge of studying evaluation theory. *New Directions for Evaluation*, 97, 57–67.
- Knaap, G. J., & Kim, T. J. (1998). Introduction: Environmental program evaluation. Framing the subject, identifying issues. In G. J. Knaap & T. J. Kim (Eds.) *Environmental program evaluation. A primer* (pp. 1–20). Urbana, USA: University of Illinois Press.
- Kraft, M. E., & Furlong, S. R. (2010). *Public policy. Politics, analysis, and alternatives*. Washington, DC: CQ Press.
- Lehtonen, M. (2005). OECD environmental performance review programme. Accountability (f)or learning? *Evaluation*, 11, 169–188.
- Lehtonen, M. (2006). Deliberative democracy, participation, and OECD peer reviews of environmental policies. *American Journal of Evaluation*, 27, 185–201.
- Lenton, T. M., Held, H., Kriegler, E., Hall, J. W., Lucht, W., Rahmstorf, S., et al. (2008). Tipping elements in the Earth's climate system. *Proceedings of the National Academy of Sciences*, 105(6), 1786–1793.
- Martinuzzi, A. (2004). Sustainable development evaluations in Europe—market analysis, meta evaluation and future challenges. *Journal of Environmental Assessment, Policy and Management*, 6(4), 411–442.
- Mickwitz, P. (2003). A framework for evaluating environmental policy instruments, context and key concepts. *Evaluation*, 9, 415–436.
- Mickwitz, P., & Birnbaum, M. (2009). Key insights for the design of environmental evaluations. In M. Binbaum & P. Mickwitz (Eds.) *Environmental program and policy evaluation. New Directions for Evaluation*, 122, 105–112.
- Moberg, F., & Galaz, V. (2005). *Resilience: going from conventional to adaptive freshwater management for human and ecosystem compatibility*. Stockholm: SIWA (Swedish Water House Policy Brief, No. 3).
- Nilsson, M., Jordan, A., Turnpenny, J., Hertin, J., Nykvist, B., & Russel, D. (2008). The use and non-use of policy appraisal tools in public policy making: An analysis of three European countries and the European Union. *Policy Sciences*, 41, 335–355.
- Owens, S., Rayner, T., & Bina, O. (2004). New agendas for appraisal: Reflections on theory, practice and research. *Environment and Planning A*, 36(11), 1943–1959.
- Pahl-Wostl, C. (2009). A conceptual framework for analysing adaptive capacity and multi-level learning processes in resource governance regimes. *Global Environmental Change*, 18, 354–365.
- Paterson, M. (2009). Post-hegemonic climate politics? *British Journal of Politics & International Relations*, 11(1), 140–158.
- Pawson, R. (2006). *Evidence based policy: A realist perspective*. London, UK: Sage.
- Pielke, R. A. (2004). What future for the policy sciences? *Policy Sciences*, 37, 209–225.
- Pielke, R. A. (2007). *The honest broker. Making sense of science in policy and politics*. Cambridge, UK: Cambridge University Press.
- Ridder, D., Mostert, E., & Wolters, H. A. (Eds.) (2005). *Learning together to manage together. Improving participation in water management*. Osnabrueck, Germany: University of Osnabrueck, USF.
- Russel, D., Haxeltine, A., Huitema, D., & Nilsson, M. (2010). Climate change appraisal in the EU: Current trends and future challenges. In M. Hulme & H. Neufeldt (Eds.) *Making climate change work for us* (pp. 31–53). Cambridge, UK: Cambridge University Press.

-
- Sanderson, I. (2002). Evaluation, policy learning and evidence-based policy making. *Public Administration*, 80(1), 1–22.
- Schreurs, M., & Tiberghien, Y. (2007). Multi-level reinforcement: Explaining EU leadership in climate change mitigation. *Global Environmental Politics*, 7(4), 19–46.
- Stufflebeam, D. (2001). The meta evaluation imperative. *American Journal of Evaluation*, 22, 183–209.
- Van de Graaf, H., & Hoppe, R. (1996). *Beleid en politiek. Een inleiding tot de beleidswetenschap en de beleidskunde*. Bussum, The Netherlands: Coutinho.
- Vedung, E. (2005). *Public policy and program evaluation* (3rd ed.) New Brunswick, USA: Transaction Publishers.
- Weiss, C. (1977). Research for policy's sake: The enlightenment function of social research. *Policy Analysis*, 3(4), 531–545.
- Wollmann, H. (2007). Policy evaluation and evaluation research. In F. Fischer, G. J. Miller, & M. S. Sidnet (Eds.) *Handbook of public policy analysis* (pp. 393–402). Boca Raton, FL: CRC Press.