Governing Policy Evaluation? Towards a new Typology

Jonas J. Schoenefeld[1] and J. Andrew Jordan

Abstract

As policy evaluation matures, thoughts are turning to its governance. However, few scholars have combined insights from the evaluation and governance literatures to shed new light on this matter. In order to address this important gap, this paper develops a new typology of ways to comprehend and perhaps ultimately govern *ex-post* policy evaluation activities. The paper then explores its validity in the context of climate policy evaluation activities, a vibrant policy area in which the demand for and practices of evaluation have grown fast, particularly in Europe. The analysis reveals that the typology usefully guides new thinking, but also highlights important gaps in our empirical knowledge of the various modes of governing policy evaluation. The paper identifies a need for a new research agenda that simultaneously develops a fuller understanding of these evaluation practices and the options for governing them.

Keywords: policy evaluation, evaluation governance, evaluation policy, polycentric governance, climate change, policy monitoring

Please note: This manuscript has been accepted for publication in the Journal Evaluation. Minor changes may still occur in the proofing process – please check the latest version for updates.

[1] Corresponding author.

Jonas Schoenefeld is Research and Teaching Fellow at the Technische Universität Darmstadt and a PhD candidate at the Tyndall Centre for Climate Change Research, School of Environmental Sciences, University of East Anglia, UK. P: 01603 593900
E: j.schoenefeld@uea.ac.uk

Andrew Jordan is a Professor of Environmental Policy at the Tyndall Centre for Climate Change Research, School of Environmental Sciences, University of East Anglia, UK. P: 01603 592552 E: A.Jordan@uea.ac.uk

Introduction

Governance scholars have spent many decades conceptualizing and empiricising the various forms of governance, whether they be hierarchical, decentralized or networked (Levi-Faur, 2012). However, somewhat surprisingly, few attempts have been made to apply these insights to the practice of *ex-post* policy evaluation. This paper follows Vedung (1997, p. 3) in defining policy evaluation as a "careful retrospective assessment of the merit, worth, and value of administration, output and outcome of government interventions, which is intended to play a role in future practical action situations."

The lack of sustained attention given to the governance of evaluation activities may stem from the fact that evaluation has thus far mainly been considered in highly specialized communities, in which (important) tasks such as developing evaluation methodologies, guidance for making value judgements, and accounting for patterns of knowledge utilization have been deemed to be especially paramount (e.g., Alkin & Christie, 2004; Patton, 1997; Pawson & Tilley, 1997; Vedung, 1997). By contrast, the various ways of organizing - or governing - evaluation have not been explicitly addressed in recent typologies and 'theories about the practice of evaluation' (Leeuw & Donaldson, 2015, p. 470). Given the growing attention to and interest in policy evaluation by policy-makers and others (e.g., EEA, 2016), coupled with a corresponding growth in investments in evaluation, growing evaluation communities, evaluation activities, and outputs (e.g., Jacob, Speer & Furubo, 2015; Furubo, Rist, & Sandahl, 2002; Mastenbroek et al., 2015; Toulemonde, 2000), now seems an opportune moment to pose important questions about the governance of evaluation itself (see, for example, Hanberger, 2012; Stame, 2006).

This paper seeks to develop a new typology in order to comprehend patterns of *ex-post* evaluation conducted by many different kinds of organizations and to provide options for perhaps ultimately governing evaluation. These two objectives will hopefully make the paper relevant not only for evaluation theorists, but also for practitioners who may wrestle with more applied questions on how to organize evaluation activities. Our typology draws on the well-known distinction between formal and informal evaluation activities (e.g., Weiss, 1993; Hildén et al., 2014). It also draws on new thinking from polycentric governance, that is, governance activities spread across multiple independent governance centres and levels (see V. Ostrom, 1999), in order to provide deeper insights into the relative merits of conceptualizing and ultimately governing evaluation in more or less hierarchical ways. It also builds on the efforts of earlier scholars, who have advanced the concept of 'evaluation policy', which refers to managing evaluations within a *single* organization (e.g., Trochim, 2009). In doing so, this paper problematizes deeper and more long-standing assumptions about what constitutes 'good' evaluation practice. Evaluation scholars have implicitly raised these questions before. Writing in the pages of this journal, Jacob, Speer and Furubo (2015) considered pluralistic evaluation systems to be more advanced than monocentric or hierarchical ones, but did not fully justify and evidence their claims. Their attempt was symptomatic of a collective failure to draw on governance theories to comprehend evaluation activities, which often bring together multiple actors and interests.

However, this is not for want of trying. For example, a decade ago Mickwitz (2006, p. 71) asked:

> Should the evaluation requirement of EU environmental policies be operationalized by the EU Commission commissioning evaluations? Or by the Council, the Parliament, the Member States or perhaps by the European Environment Agency? Should there be some institutions with specific capacities to conduct evaluations, as in some countries and sectors and what are the pros and cons of different structures?

These remain highly pertinent questions, not least in the context of the transparency provisions in the Paris Agreement on climate change (Schoenefeld et al., 2016; UNFCCC, 2015). Whilst there is some knowledge about the political struggles that have emerged around the allocation of evaluation roles to particular institutions (Martens, 2010), in the case of climate and environmental policies, evaluation and governance scholars have made notably little progress in jointly conceptualizing and perhaps ultimately governing evaluation practices.

This paper draws on the empirical case of environment and particularly climate change policy evaluation in the European Union (EU) in order to test the new typology. Although these were not the first sectors in which EU policy evaluation developed (see Stame, 2003; 2006; Mickwitz, 2006; Crabbé & Leroy, 2008; Toulemonde, 2000), recent studies have revealed them to be especially dynamic sites of *ex-post* evaluation (e.g., EEA, 2016; Haug et al., 2010; Huitema et al., 2011; Hildén et al., 2014; Hildén, 2011; Mickwitz, 2013). The European Environment Agency recently argued that “[t]he evaluation of environment and climate policies is, today, a well-established discipline” (EEA, 2016, p. 4). Correspondingly, Huitema and colleagues (2011) found 259 *ex-post* climate policy evaluation documents between 1998 and 2007 (see also Haug et al., 2010).

Similarly, at the time of this writing, a German database of energy policy ‘studies’ lists 243 documents related to ‘climate change’.[2]

Policy evaluation in the EU receives considerable attention from high-level policy actors (see Mickwitz, 2013), such as the European Commission (e.g., Mastenbroek et al., 2015), the European Environment Agency (EEA, 2016; Martens, 2010; Hildén et al., 2014), the European Court of Auditors (Stephenson, 2015) and the Member States (Furubo, Rist & Sandahl, 2002). Stern (2009) has estimated that the European institutions spend approximately 45 million Euros per year on evaluation; Hojlund (2015) has calculated that the European Commission alone employs 140 staff to manage it. However, government-driven climate policy evaluation activities remain highly differentiated across the EU, where countries such as Germany and the UK have significant evaluation capacities, but southern and new member states exhibit much lower activity levels (AEA, ECOFYS, Fraunhofer, & ICCS, 2009, p. 33; Jacob, Speer, & Furubo, 2015). Outside government, there are also many other actors, such as environmental groups, that should be accounted for (e.g., Haug et al., 2010; Hildén et al., 2014; Huitema et al., 2011; Mickwitz, 2013). Growing attention to evaluation and a wide variety of actors, practices, and evaluation outputs make climate policy a suitable area in which to explore the everyday practices and governance of evaluation within a single political system, namely the EU (see also Stern, 2009; Jacob, Speer & Furubo, 2015).

The remainder of this paper unfolds as follows: in the second section, the paper draws on evaluation literatures to review the important distinction between formal (i.e. state-led) and informal (i.e. society-led) policy evaluation actors. The third section draws

[2] http://www.forschungsradar.de/studiendatenbank.html

on (polycentric) governance literatures to develop a second continuum ranging from hierarchical to polycentric ways of conceptualizing evaluation. These two continua inform a new typology, which the paper introduces in the fourth section, and then discusses in detail with a view to understanding on-going climate policy evaluation activities in the EU, as documented in the existing literature. The fifth section discusses our results and reflects on the fruitfulness of the typology as well as opportunities for new research.

Formal or informal evaluation?

For analytical purposes, evaluation scholars have found it useful to distinguish between formal (i.e., government-driven) and informal (i.e., society-driven) modes of evaluation. In a ground-breaking article, Weiss (1993) distinguished between ‘inside evaluation’ conducted by people ‘inside’ government, and ‘outside’ evaluation by actors not linked with government (see also Conley-Tyler, 2005). Other researchers have developed the related notions of ‘formal’ versus ‘informal’ evaluation in the EU (Hildén et al., 2014; Huitema et al., 2011). Hildén and colleagues (2014) define formal evaluation as ‘state-led’ and informal evaluation as ‘evaluation activities by non-state actors’ (p. 885). Crucial for this paper is the fact that each mode comes with a set of potential strengths and weaknesses, which the following section considers.

Formal evaluation

In-house or formal evaluators may have intimate knowledge of policy processes and the circumstances under which a particular policy emerged, which may in turn make the evaluation more attuned to these and other contextual variables (Weiss, 1993;

Toulemonde, 2000). Careful attention to such factors may eventually facilitate the uptake of evaluation knowledge later on in the policy process (Weiss, 1993), an issue that remains a core concern amongst scholars evaluation (see, for example, Albaek, 1995; Chelimsky, 2006; Fischer, 2006; Hertting & Vedung, 2012; Patton, 1997; Toulemonde, 2000). Furthermore, if governmental actors fund evaluation activities, they may be under considerable pressure to act upon related findings, or respond to them publically.

Conversely, formal evaluation also has well known weaknesses. Evaluation findings by governmental evaluation actors may be less critical of a given policy and its outcomes than evaluative knowledge generated by non-state actors (Weiss, 1993). Political pressures to 'look good' and avoid negative evaluations may be immense and thus inhibit formal actors from being too critical. Or formal evaluators may knowingly or unknowingly seek evidence in order to support their pre-existing hypotheses or views on a policy by way of a 'confirmation bias' (see Nickerson, 1998). And if unfavourable evaluation results do emerge, governmental actors may have an incentive to suppress them or not draw attention to them if they are published – an important issue, given that evaluation can also inform public debates (Chelimsky, 2006). Evaluation scholars have long argued that it is important to protect the independence of evaluators within government (see Chelimsky, 2009).

Such political pressures also emerge when governmental actors commission organizations outside government to conduct evaluations. For example, evaluation criteria that policy-makers focus on may differ significantly from criteria that those who are subjected to a policy may perceive as adequate (Weiss, 1993; Majone, 1989, p. 168). Furthermore, the commissioning process may generate principle-agent relationships, which may trigger political struggles around policy evaluation. For example, a recent survey of evaluators in the UK revealed how civil servants have attempted to directly influence the outcome of evaluations (Hayward et al., 2013); another approach involves trying to frame evaluation findings in a more positive light (Weiss, 1993). According to Hayward and colleagues (2013), UK civil servants have used a range of strategies to achieve this, including controlling the research questions addressed by evaluations, or by enacting budgetary-turned-methodological constraints (such as not funding a control group). However, interference at earlier stages in the evaluation process appeared more popular than influence at later stages (Hayward et al., 2013). More recent research has confirmed the existence of these dynamics in other countries (see Pleger & Sager, 2016). Another tactic by those who see evaluation as a useless 'bureaucratic burden' has been to allow evaluators very little time to conduct evaluations, leading to superficial results (Toulemonde, 2000). In a similar vein, Stame (2004, p. 504) concluded that: "[e]valuation in Europe suffers from being too constrained by the demands of those who commission evaluations, and by the regulations that are put in place." In sum, the closeness of formal evaluators to the policy-process may make evaluations more realistic and facilitate uptake. However, the findings are likely to be less radical - indeed civil servants may have incentives to exert continuing influence over evaluators.

Informal evaluation

By contrast, evaluations performed by non-state actors may take a more critical look at policies (Weiss, 1993). This is because informal evaluators may have fewer incentives to 'look good' or potentially downplay negative aspects of a policy. In fact, informal evaluators—or their funders—may conduct evaluations precisely to expose the shortcomings of policies in order to pressurise policy-makers to respond. The latter may generate substantial incentives to bring evaluation results into public discussions (see Chelimsky, 2006). By the same token, informal evaluations are more likely to be 'reflexive' (see Fischer, 2006), meaning that they may have more room to critically reflect on extant policy objectives. Informal evaluation may also employ a greater range of criteria (Mickwitz, 2013) in order to pay more attention to policy side effects (see Vedung, 2013) that may not feature in 'distance to target' policy evaluation exercises conducted by formal evaluators (Hildén et al., 2014). By drawing on non-governmental resources, informal evaluation may also be less affected by electoral and budgetary cycles, as well as shifting political priorities within government. Crucially, they may in principle emerge in the absence of central coordination and stimulation. Of course the true extent to which this happens remains an (open) empirical question.

Informal evaluation activities may also exhibit a range of potential weaknesses. Informal evaluators may not have detailed 'inside' knowledge and may thus overlook key aspects of a policy or be oblivious to the (political) process through which a policy first emerged (Weiss, 1993). More critical evaluation results may also prove much less palatable for policy-makers, thus leading to a lower uptake of evaluation knowledge or potentially even outright resistance. Informal actors may simply struggle to fund costly and rigorous evaluation exercises. For example, Löwenbein (2008) estimated a cost of about one hundred thousand Euros per evaluation of a German structural fund project. In comparison to many governmental actors, all but the most well-funded non-governmental actors may struggle to muster such resources (see Greenwood, 2011, p. 136-141). The aforementioned 45 million Euros spent by the EU institutions on evaluation every year

(Stern, 2009), for example, amounts to more than ten times the total budget of the WWF's European Policy Office (which is one of the best-funded in Brussels).[3] By the same token, informal actors may also have vested interests that diverge from the interests of the wider public. If they have a more or less pre-determined view of preferred policy outcomes, it too can lead to 'confirmation bias'. Substantial funding (such as that available to the fossil fuel industries in the case of climate change) may generate a situation where 'money evaluates', rather than evaluators, compromising idealised visions of 'systematic' or 'pluralistic' evaluation (see Jacob, Speer & Furubo, 2015). In short, formal and informal evaluators and/or their funders may have considerable incentives to use policy evaluation as another weapon in policy battles. It is thus an open question as to whether evaluation activities really emerge organically from the bottom up - a point to which the paper shall return.

Summary

With a view to individual policy evaluations, there are thus numerous considerations that flow from the involvement of formal and informal actors in evaluation (see Table 1). There is, in short, no approach which is obviously 'better' – each has strengths and weaknesses. Drawing on the work of polycentric governance scholars, it is likely that the efficacy of different approaches depends on their 'fit' with overall socio-political circumstances (see E. Ostrom, 1990). However, if evaluation is to contribute to a better understanding of socio-environmental systems, it cannot be limited to evaluating single policies. Thus a key question is how to govern the evaluation of multiple policies

[3] http://ec.europa.eu/transparencyregister/public/consultation/displaylobbyist.do?id=1414929419-24

across many scales and sectors of governance, such as exists in the EU. In the EU, governors have deployed well over one thousand separate policies to address climate change (Schoenefeld et al., 2016). The following section unpacks the strengths and weaknesses of two potential forms of organising evaluation activities, namely hierarchical and polycentric.

Insert Table 1 here.

Hierarchical or polycentric evaluation?

So far, the discussion has more or less assumed one level of governance and hence level of evaluation activities. However, in line with many other governance processes, policy evaluation has evolved into an increasingly multi-level affair (see Hooghe & Marks, 2010; Stame, 2008). In order to capture this aspect, we differentiate between two ways of governing: hierarchical (or top-down evaluation) conducted and coordinated by a central actor; and more decentralised (or polycentric) potentially conducted by 'self-organising' evaluators. This section disentangles the strengths and weaknesses of each approach.

Hierarchical evaluation

In principle, evaluation activities organised by a single actor – which could be but need not necessarily be a state - exhibit several strengths. A single evaluator (or institution) may be able to set common evaluation standards and thereby make the results across multiple policies more comparable, which is a key concern, especially regarding climate policy (e.g., Aldy & Pizer, 2015; Aldy, Pizer & Akimoto, 2016; Feldman & Wilt, 1996; Purdon, 2015; Schoenefeld et al., 2016). A single evaluator may also have greater resources than multiple smaller actors, which may translate into stronger evaluation capacities. Hierarchical evaluation can also facilitate the coordination of evaluation activities in order to avoid duplication of costly analysis, and provide one central location from which evaluative knowledge diffuses. Crucially, a central actor may be able to exert the 'political pressure' needed to foster effective coordination of evaluation (for related arguments on policy coordination, see Jordan & Schout, 2006, p. 271; Peters, 1998). For example, De Burca and colleagues (2014) suggest that in a context of 'global experimentalist governance', "a new kind of centre [is required], pooling information and organizing peer evaluation of it, and on occasion responding to (or invoking the threat of) a penalty default" (p. 478-79).

By the same token, hierarchically-organised evaluation may also suffer from several weaknesses. For example, streamlined standards may prove insensitive to contextual effects, such as (un)intended side effects, which can be crucial factors in judging the success and/or failure of particular policies (e.g., E. Ostrom, 2010; Thompson, Rausch, Saari, & Selin, 2014). For example, the Intergovernmental Panel on Climate Change (IPCC), the main provider of authoritative scientific advice to governments, has highlighted many potentially beneficial side effects of climate policy. But it remains questionable whether standardising evaluation adequately captures them all (Somanathan et al., 2014). Furthermore, centrally organised evaluation can be perceived as a way of 'policing' policy performance aimed at control, which could potentially provoke resistance from lower-level actors (see Stame, 2008; Schoenefeld et al., 2016). If such resistance morphs into minimal or even no cooperation, the quality of evaluation may suffer since evaluators may for example depend on information from the actors who are being evaluated (Chelimsky, 2006). Hierarchically organised evaluation may also suffer from higher risks of systemic failure, given that it may be difficult to change approaches and standards if they prove unsuitable (for the general argument, see E. Ostrom, 2010). Or it may simply miss more innovative policies, which have not yet been incorporated into formal monitoring and evaluation systems. Finally, in many settings, such as the United Nations Framework Convention on Climate Change (UNFCCC), a single, hierarchical evaluation actor may simply be politically infeasible in the short term.

Polycentric evaluation

If, as Carlsson (2000) suggests, complex policy environments benefit from more bottom-up evaluation methodologies, then what about the actors that function at ‘the bottom’? The strengths and weaknesses of polycentric evaluation are the inverse of hierarchical evaluation. Polycentric evaluation in principle exhibits more flexibility and sensitivity to context as evaluators can adjust evaluation criteria to local circumstances. According to classic arguments by polycentric governance scholars, this flexibility translates into lower risk of systemic failure, because evaluators can address problems locally, and failure in one part of the evaluation system does not necessarily generate systemic failures (see E. Ostrom, 2010). Furthermore, more local evaluation may lead to more local ownership of evaluation results, and reduce perceptions that they are being ‘policed’ from the top. Local ownership could also lead to a greater uptake of evaluation knowledge. Finally, local actors may face incentives to spread their evaluation findings to others and even profit from them. Particularly early adopters may want to share their experiences with others and thus engage in consulting activities. Crucially, polycentric evaluation does not rely on a single evaluation actor and thus proves more suitable to situations in which a dominant governor does not exist, as is currently the case with many transnational efforts to address climate change (Chan et al., 2016; Ostrom, 2014; Widerberg & Stripple, 2016).

Polycentric evaluation may, however, also exhibit a range of weaknesses in the absence of a central coordinator. Multiple, localised evaluation standards may stifle the ability to compare evaluation results across multiple policies and draw cumulative conclusions. Whether or not conversion in evaluation methods and standards happens without coordination remains an open question. Relatedly, if evaluation approaches and standards change frequently, it can become difficult if not impossible to track policy development over time. This may be particularly problematic for policies with long life spans, such as those in the climate change area. Given the cost of some evaluation activities (see, for example, Löwenbein, 2008), evaluation and evaluative knowledge may not be freely available, and thus be subject to of collective action dilemmas identified in earlier literatures (e.g., Ostrom, 1990). Furthermore, while early adopters of evaluation may be keen to share their findings with others, the incentive to share lessons from failed attempts may be much lower or even non-existent. Such experiences are just as important as those related to success, because they can prevent others from similar mistakes. But who will communicate those failures and related insights?

Not all agree, however. A key insight from the polycentric governance approach (see E. Ostrom, 1990; V. Ostrom, 1999) is that local actors may enjoy considerable selfgoverning capabilities. In recent decades, empirical evidence has emerged which emphasises how local actors manage to build enduring institutional systems to monitor and govern their local resource use (E. Ostrom, 1990). They may be in a better position to govern evaluation in ways that better fit local contexts (E. Ostrom, 1990). In such circumstances, panaceas (i.e. 'one-size-fits-all' approaches based on abstract reasoning from first principles) are less likely to work (E. Ostrom, Janssen, & Anderies, 2007). But do these insights hold for all evaluation activities? While Elinor Ostrom (2005, p. 280) argued that local actors may in principle have the capacity to pool resources in order to conduct evaluations, it remains unclear whether and to what extent the self-organisation of evaluation is indeed an empirical reality.

Summary

Taken together, governing evaluation activities hierarchically or polycentrically comes with a range of potential strengths and weaknesses, which Table 2 summarises.

Insert Table 2 here.

Governing Evaluation

So far, our discussion has revealed that there is no inherently 'better' way of governing evaluation. In fact combining the two dimensions produces a 2×2 typology which opens up a range of potential combinations of strengths and weaknesses. Figure 1 details four key modes of governing evaluation given that both dimensions (formal/informal, hierarchical/polycentric matter. This section asks to what extent it is possible to use this typology to comprehend on-going evaluation activities, and thus potentially explore new ways of governing them in the future. Importantly, our typology includes both actor types (the two continua) and the standards and methods used by these actors (in each of the quadrants). This section draws on the existing literature in order to explore the extent to which these four modes can be detected in the EU. In doing so, it seeks to provide a basic 'plausibility probe' (Eckstein, 2000) of our typology to assess its ability to organize extant empirical knowledge and comprehend potential new ways to govern evaluation.

Insert Figure 1 here.

Formal common standards and methods

The top left corner of Figure 1 harbours formal common standards and methods enacted by governmental actors in a hierarchical fashion. At the EU level, considerable efforts have been made to harmonise and institutionalise evaluation practices. The European Commission – arguably one of the most important advocates and exponents of evaluation in the EU (Mickwitz, 2013) – has published a series of communications which have sought to encourage and systematize evaluation (European Commission, 1996; 2007; 2013). A 2007 communication includes a set of evaluation standards in order to streamline evaluation, as older standards were perceived to be unable to produce high-quality evaluations. However (and in line with our conceptualisation), the limitations of its approach quickly became apparent to scholars. For example, Mickwitz (2013) pointed out that the 2007 evaluation standards did not include side effects. Furthermore, the 2007 communication clearly highlighted the tension between hierarchical control and evaluator independence by asserting that "evaluators must be free to present their results without compromise or interference, although they should take account of the steering group's comments on evaluation quality and accuracy" (European Commission, 2007, p. 23). Evidence from the UK and other countries (see Hayward et al., 2013; Pleger & Sager, 2016) raises doubts about the feasibility of maintaining such a stance. Stern (2009) too wrote that at EU level "there is a widespread perception in the evaluation community that independence is not always highly valued" (p. 72).

Hierarchical forms of evaluation can also trigger considerable resistance from lower levels (Stame, 2008), a tendency which can certainly be observed in EU climate policy. As signatories of the UNFCCC, from 1993 the EU member states implemented greenhouse gas emission and eventually policy reporting requirements through a bottom-up 'Monitoring Mechanism' (Haigh, 1996; Hyvarinen, 1999). Although many EU member states have signed up to the need for more *ex-post* evaluation in principle, in the area of climate change they were reluctant to centralise climate policy monitoring in the European Commission or the European Environment Agency (Hildén et al., 2014; Schoenefeld et al., 2016), even though greater standardisation had been repeatedly recommended by researchers (Mela & Hildén, 2012). So although the Monitoring Mechanism has been revised twice (in 2004 and 2013 respectively)[4], "[…] less than 10%

[4] http://ec.europa.eu/clima/policies/g-gas/monitoring/index_en.htm

of the entries in the 2011 reporting cycle included quantitative data based on *ex post* evaluations […]" (Hildén et al., 2014, p. 898). In other words, most of the monitoring generates *ex-ante* predictions of what member states *hope* their policies will deliver (Hildén et al., 2014; Schoenefeld et al., 2016). Furthermore, EU member states are reluctant to allow the Commission to elicit more detailed, policy-specific data; many prefer to report on the effectiveness of 'bundles' of policies rather than individually (Hildén et al., 2014). Withholding information on individual policy instruments could be one strategy to mask the ineffectiveness of particular policy instruments in order to protect particular 'instrument constituencies' (Voß & Simons, 2014; see also Kerr, 2007). Taken together, while the EU has regularly estimated the impacts of its climate policies since the early 1990s, little of this activity draws on *ex-post* data; in fact there is considerable political resistance to giving EU-level actors more control (Schoenefeld et al., 2016).

Meanwhile, the Commission has also tried to harmonise climate policy evaluation among member states. In 2009 and 2012, it commissioned two studies with a view to streamlining standards (AEA et al., 2009; Öko-Institut, Cambridge Economics, AMEC, Harmelink Consulting, & TNO, 2012). However, these studies were rather critical of harmonisation. For example, the Öko-Institut (2012, p. iv) identified various obstacles to methodological streamlining, concluding that there was no 'one size fits all' solution to evaluation (see also Toulemonde, 2000). On-going technical disputes over measuring the greenhouse gas content of certain sources have further undermined efforts to promote greater centralisation. A controversy between the EU and Canada over the cumulative greenhouse gas emissions from the production of oil from tar sands showed that even relatively technical estimates about the greenhouse gas content of certain fuels can at times become intensely political (see Neslen, 2011). It is also worth noting that the drivers of centralisation vary considerably at the EU member state level: Mela and Hildén (2012) found that the UK had issued considerably more guidance than other EU member states.

Taken together, enacting common evaluation standards and methods has created intense political conflict (Schoenefeld et al., 2016). However, EU member states have often attempted to relegate this to the domain of the 'unpolitical' (see Hildén et al., 2014). Thus, while governmental actors appear to be investing in evaluation (see above), the other putative strengths of more top-down and hierarchical approaches have not yet materialized, namely with a view to generating common evaluation and monitoring standards.

Informal common standards and methods

The right top quadrant of Figure 1 contains common standards and methods, enacted by societal actors, who also engage in evaluation activities. One prominent example is the 'European Environment Evaluators Network' (EEEN), which belongs to the much larger, international 'Environmental Evaluators Network'. It aims to bring environmental evaluators together in order to facilitate knowledge exchange.[5] However, while a range of formal and informal actors are now involved, the original impetus for this network appears to be formal, driven by the United States (US) Environmental Protection Agency and the US-based National Fish and Wildlife Foundation, a non-profit

[5] http://www.environmentalevaluators.net/purpose/

grant-giving conservation organisation overseen by the US federal government[6]. Rather than proposing evaluation standards, these organisations endeavour to spread evaluation knowledge and 'best' practice. Another, more climate policy-focused example is Climate-Eval[7], a "community of practice set up by the IEO [Independent Evaluation Office of the UN's Global Environmental Facility] with donor support in 2008…" (Uitto, 2016, p. 111). It maintains an email list, publishes guides on policy evaluation (see Woerlen, 2013), and maintains a database of evaluation studies. Similar to the EEEN, this is an international 'community of practice' that works through informal, peer to peer knowledge exchange and learning. This approach has also manifested itself in the emergence of evaluation societies in numerous countries and at EU level (Jacob et al., 2015).

Informal actors have also become involved in organising evaluation knowledge. In contrast to academia, where knowledge management and database systems are relatively advanced, to date there are no integrated evaluation knowledge management systems in evaluation (however, sub-systems do exist in some fields such as development aid—see Liverani & Lundgren, 2007). In the area of climate change, several actors have attempted to create databases of evaluation documents. Important (but still limited) examples include an online database managed by the European University Institute[8], one in Germany focusing on renewable energy policy evaluations[9], another managed by the

[6] http://www.nfwf.org/whoweare/Pages/home.aspx#.VOXD_HbrHII
[7] https://www.climate-eval.org/about
[8] https://cprubibliography.wordpress.com/
[9] http://www.forschungsradar.de

Climate-Eval initiative[10], and the Architectures of Evaluation approach[11] pursued by the US Environmental Protection Agency and the EEA. Crucially, however, formal, governmental actors support the German and the Climate-Eval databases.

Another, and arguably different approach to standardising climate evaluation is to certify evaluators. For example, Thomas Dreesen suggested the possibility of creating an organisation of ‘chartered’ energy efficiency evaluators at the 2014 International Energy Program Evaluation Conference in Berlin (see Cooney, Dreesen, Lees, & Titus, 2014), an idea that is already being practised by the Japanese Evaluation Society on general policy evaluation (Jacob et al., 2015). Generally, this idea appears to be picking up steam in the evaluation literature (see McDavid & Huse, 2015). The idea is to create a structure similar to that of chartered accountants. Such proposals have emerged because of a growing awareness that evaluation is simply too context-specific for standardised methods. Hence, the idea is to generate a group of certified professionals to conduct evaluations.

Taken together, there are numerous informal actors that have tried to standardise evaluation practices, usually through networks, knowledge exchange activities and professional accreditation. While these actors have created some evaluation databases, in the area of climate policy they have so far not been able to produce common evaluation standards and/or metrics. Crucially, and somewhat at odds with the expectations of polycentric governance theorists (see E. Ostrom, 2005), a significant impetus for these initiatives has come from governmental actors, or at least actors that receive considerable

[10] https://www.climate-eval.org/eLibrary

[11] http://www.environmentalevaluators.net/archee/

central governmental support. This support has in turn enabled other organisations to join. In other words, governmental actors are by no means the only ones driving evaluation, given that informal actors often collaborate to work towards more cohesion in climate policy evaluation.

Commonly negotiated standards and methods

At the EU level, the European Environment Agency (EEA) is at the centre of many networks of actors who shape climate policy evaluation standards. As Martens (2010) explains, those that established the EEA in 1991 disagreed on its role: while the Commission and a number of Member States wanted it to generate environmental data, the European Parliament envisioned that it would adopt an independent and/or policy scrutinizing role (see also Waterton & Wynne, 2004). While these disagreements were eventually buried in ambiguous language in the regulation that established the EEA, the politics never entirely disappeared. Particularly in the first decade of its existence, tensions emerged between the EEA and the Commission's Directorate–General (DG) for the Environment, with the former seeking a stronger policy-analysis role, and the latter preferring more data collection (Martens, 2010). To this day, the initially strong emphasis on data collection remains (Martens, 2010; Mickwitz, 2013), but the EEA has started to indicate a willingness to engage in more policy evaluation (EEA, 2016).

The EEA plays a key role on climate policy evaluation in the EU because it operates the EU's Monitoring Mechanism for greenhouse gases and, increasingly, policies and measures to reduce them (EEA, 2016). Clearly, the revision of the Monitoring Mechanism and the significant concessions made by the Commission in this process (see above) can be understood as ultimately producing a set of 'negotiated' evaluation standards (Hildén et al., 2014). For example, in a public consultation on plans to revise the mechanism, many respondents voiced their dissatisfaction with the existing situation.[12] The negotiations took place between the national level (EU member states) and the EU level (Commission), but also included the European Parliament, which had to formally sign off the new regulation (Schoenefeld et al., 2016). These negotiations were arguably conducted in the 'shadow of hierarchy' (see Börzel & Risse, 2010), because EU member states held the upper hand. They knew that a very weak revision of the Monitoring Mechanism could result in ever-more centralised climate policy evaluation later on. It remains to be seen whether the EEA will exploit any available leeway to drive monitoring in the direction of policy evaluation.

However, it remains an open question whether these negotiated standards are sufficient: scholars have recently raised doubts about the validity of the data provided through the Monitoring Mechanism (Hildén et al., 2014; Schoenefeld et al., 2016). More broadly, Aldy (2014) has highlighted that the policy monitoring standards negotiated under the UNFCCC, which underwrite the EUs monitoring mechanism, are insufficient to track climate policy over time. Taken together, it remains unclear whether independent actors such as the EEA can negotiate and use standards that are sufficient to compare policy over time and highlight potential shortcomings, particularly given the strongly political nature of the negotiations. Climate policy evaluation in the EU thus reveals some scope for, but also very real political limits to, changing the balance between evaluation and monitoring (Schoenefeld et al., 2016).

[12] http://ec.europa.eu/clima/consultations/docs/0008/results_en.pdf

A la carte standards and methods

Given many gaps in empirical evidence, scholars know much less about the right bottom quadrant in Figure 1. Existing literatures highlight the great variety of actors involved in evaluation across the EU. As Versluis and colleagues (2011, p. 224) write: "[…] EU evaluation culture is political and pluralistic, characterized by a variety of organizations willing to pay significant sums of money to finance research that may produce data in support of their political views." Evidence from the only available large-scale meta-analysis of climate policy evaluation documents suggests that a range of formal and informal actors evaluate climate policies across the EU (Haug et al., 2010; Huitema et al., 2011). The European Commission is one of the most active producers of evaluation knowledge across sectors (Mastenbroek et al., 2015). Hildén (2014) writes that with regard to the landmark EU emissions trading scheme, informal evaluators, such as the Union of the Electricity Industry, have commissioned many evaluations. In addition, informal academic evaluators have also become ever more important (Hildén, 2014).

While Elinor Ostrom (2005, p. 283) argued that smaller organisations can in principle pool resources in order to conduct evaluations, current literatures suggest that the extent this is happening in EU climate policy is rather limited, although there are some notable exceptions such as the Climate Action Tracker (see Fransen & Cronin, 2013, for a review). When this happens, it can also generate significant benefits for local organizations: for example, by turning '100% renewable', the village of Feldheim in Germany has attracted worldwide attention (Ratzesberger, 2014).

In the meta study conducted by Huitema and colleagues (2011), less than 10 of the 259 evaluations analysed were conducted by industry or trade associations; indeed, the number of evaluations done by non-governmental organisations was less than 20 between 1998 and 2007. Thus, current data cast doubt on whether climate policy evaluation is likely to self-organise in the way that Ostrom (2005) suggested. As far as evaluation standards are concerned, Huitema and colleagues (2011) found that goal attainment and effectiveness were most widely used, as were a range of other criteria. This rather limited set of frequently used criteria could allow for some comparability in determining evaluation results. In another, smaller meta-analysis of climate policy evaluation studies, Mela and Hildén (2012) found that cost effectiveness was a more commonly used criterion, but concluded that climate policy evaluation practice tends to be very heterogeneous across the EU. Crucially, significant doubts remained as to whether informal evaluation can fill gaps left by formal evaluation actors, particularly with a view to critically reflecting on extant policy goals (Huitema et al., 2011; see also Fischer, 2006). Furthermore, there is currently no database or central repository of informal climate policy evaluation documents. In fact aside from Huitema et al. (2011), nobody has collected - let alone analysed - informal policy evaluation practices.

Discussion and conclusions

This paper started from the premise that there is much to be gained from bringing together insights from governance and evaluation to reflect upon the governance of policy evaluation itself. To test this out, this paper combined insights from evaluation studies on formal and informal evaluation with governance theories on hierarchical and polycentric governance to generate a novel typology. The paper thus makes a key theoretical contribution towards engaging with governance theory in order to comprehend and perhaps ultimately govern patterns of evaluation conducted by many different kinds of actors—an approach which leading evaluation scholars have long called for (e.g., Stame, 2006; 2008) and which goes well beyond existing efforts to establish common ‘evaluation policies’ within single organizations (see Trochim, 2009).

The paper then subjected the typology to a ‘plausibility probe’ (Eckstein, 2000) using the case of climate policy, a very dynamic policy area in the EU in which many efforts have been made to engage in monitoring and evaluation (EEA, 2016). Marshalling the currently fragmented and partial stock of existing empirical material helped to fuller understand and make sense of climate policy evaluation activities in the EU, and drew attention to what is at stake, both theoretically and empirically. By doing so, the paper demonstrates that combining theoretical insights from governance and evaluation literatures generates a number of pertinent research questions that have received too little attention thus far. More precisely, each quadrant in our typology contains a unique combination of strengths and weaknesses of different modes of evaluation practice and governance—with no obvious indication of which governance mode is ‘better’. The answer to the latter question will likely depend on the substantial policy field, as well as actor preferences on what evaluation is expected to achieve. For example, the analysis shows that formal, state-led evaluation does not necessarily have to be hierarchical, as governmental evaluators may work at various, decentralised levels (such as in federal systems).

However, from what patchy empirical evidence exists in the realm of climate policy, it appears that the strengths of each quadrant have at best only partially materialized, and the weaknesses remain ever-present across the typology. This state of affairs highlights the need for more targeted forms of data-driven analysis on evaluation governance, including in a range of other policy sectors that this paper could not consider, but where this typology may also be usefully applied, and which therefore constitutes an important venue for future research. Many of the absolutely critical dilemmas that emerge from our typology – such as the independence of evaluators, the publication of evaluation results and the ability of different governance centres to learn from one another – are difficult to resolve given the paucity of empirical evidence.

Shedding more light on these dynamics might also be useful for those seeking to govern evaluation activities. The availability of many different governance options, which Figure 1 sought to distil, means that important questions are at stake. Who gets to decide what evaluation governance mode is most suitable? Whose preferred criteria are most relevant for what is perceived as a ‘functioning’ evaluation system? As the paper detailed above, new evaluation activities will themselves generate winners and losers (e.g., those who receive funds to produce evaluations, and those who do not, or those whose funds are cut because they are being diverted towards evaluation activities), and thus be the focus of political struggles over resource distribution, access, legitimacy and others. Future research should focus on exploring these dynamics in much greater detail than this article has been able to accomplish.

Future research should also consider the interactions between different modes of organisation and/or explore the extent to which reality may simultaneously exhibit aspects of some or even all four quadrants. For example, it is known that on occasions governmental organizations do allow evaluators considerable independence thus ameliorating some of the potential drawbacks of formal evaluation (Chelimsky, 2006; 2009; Uitto, 2016). This is especially relevant giving the growing interest in other forms of co-governance (e.g., Tosun, Koos & Shore, 2016). Our analysis reveals that, for example, formal-informal interactions appear relatively common, but much more detailed empirical investigation is needed to understand how they perform. As a first step, it would be helpful to build a more comprehensive database of formal and informal (climate) policy evaluation documents and then use evaluation and governance theory to analyse it, as well as interview actors to probe some of the underlying political and process-based aspects of evaluation. Work along these lines could provide a much needed opportunity to investigate the actor categories identified in this paper more thoroughly. While our probe shows that the categories may blur somewhat in practice (see also Guha-Khasnobis, Kanbur, & Ostrom, 2006), they provide useful theoretical yardsticks to anchor a discussion about different approaches to policy evaluation. For example, many 'informal' civil society organisations receive substantial EU funds (Greenwood, 2011). Such explorations could help shed light on the extent to which governors are able to choose freely from the menu of governance modes depicted in our typology. At present, some do appear to require much more self-organising capacity and coordination than others, the potential sources of which are still far from clear.

Finally, in addition to the public policy focus that this paper adopted, *non*-policy approaches, such as are appearing in the so-called transnational governance realms, are becoming more important in climate domain (e.g., Chan et al., 2016; Bulkeley et al., 2014). For example, the international Covenant of Mayors, which addresses energy efficiency governance in cities, highlights the need to monitor and evaluate these and other softer, network-based form of governing.[13] Similarly, the Compact of Mayors to

[13] http://www.covenantofmayors.eu/actions/monitoring-action-plans_en.html

address climate change has established systems to track progress towards publically stated targets.[14] But little has been done to monitor and evaluate such initiatives (see for example Chan et al., 2016; Widerberg & Stripple, 2016). Indeed, scholars have barely begun to map their existence.

Such research could help to refine policy evaluation systems not only in the EU (see EEA, 2016), but also with respect to other actors who may wish to evaluate. This is particularly pertinent in the wake of the Paris Agreement (UNFCCC, 2015), whose backbone is a new, 5-year review and transparency mechanism. This mechanism strongly links with energetic and fast-moving debates on developing successful climate policy monitoring and evaluation arrangements (Aldy, 2014; Aldy & Pizer, 2015; Aldy, Pizer & Akimoto, 2016; Feldman & Wilt, 1996; Jordan et al., 2015; Fransen & Cronin, 2013; Schoenefeld et al., 2016). Over time, national emission reduction targets are expected to become more stringent as the pledge and review mechanism kicks in (if not the probability of achieving the meta policy goal of keeping warming within two degrees Celsius will be extremely low). But it will become much harder for countries to fulfil their targets at reasonable cost in the absence of sound systems for evaluating individual policies and measures. In the past, the EU has had less need to evaluate its own policies, given that its climate targets have been comfortably attained through ‘non-climate policy’ effects, such as the ‘dash for gas’ in the UK or economic restructuring following reunification in Germany (Jordan, Huitema, Van Asselt, Rayner, & Berkhout, 2010). But going forwards, governors are likely to face much more pressing demands to know which of their various policies are really performing (and on what criteria). Well-developed

[14] http://www.iclei.org/details/article/global-mayors-compact-shows-unity-and-ambition-to-tackle-climate-change-1.html

evaluation systems could in principle furnish such knowledge, which will build trust within and, crucially, between countries. The EU has historically occupied a leading role in designing ways to govern evaluation in the UN climate agreement (Yamin & Depledge, 2004, p. 327). The Paris Agreement arguably provides a new 'opportunity structure' for many more actors to become involved in climate governance (Tosun & Schoenefeld, 2016), not least through the means of evaluation.

But it is important to recognize that, however organized or governed, evaluation does not exist in a vacuum. There are likely to be interactions between the way policy evaluation functions and is organized and the structure and functioning of wider governance systems. Furthermore, there are related forms of evaluation, such as *ex-ante* impact assessment, that could be subjected to similar questions. Much like the evaluation practices themselves, the forms of evaluation governance remain very much in flux.

References

AEA, ECOFYS, Fraunhofer, & ICCS. (2009). *Quantification of the effects on greenhouse gas emissions of policies and measures: Final report.* ENV.C.1/SER/2007/0019. Brussels: European Commission.

Albaek, E. (1995). Between knowledge and power: utilization of social science in public policy making. *Policy Sciences*, *28*(1), 79-100.

Aldy, J. E. (2014). The crucial role of policy surveillance in international climate policy. *Climatic Change*, *126*(3-4), 279-292.

Aldy, J. E., & Pizer, W. A. (2015). Alternative metrics for comparing domestic climate change mitigation efforts and the emerging international climate policy architecture. *Review of Environmental Economics and Policy*. DOI: 10.1093/reep/rev013

Aldy, J. E., Pizer, W. A., & Akimoto, K. (2016). Comparing emissions mitigation efforts across countries. *Climate Policy* (online first). DOI:10.1080/14693062.2015.1119098

Alkin, M. C., & Christie, C. A. (2004). An evaluation theory tree. *Evaluation roots: Tracing theorists' views and influences*, 12-65.

Börzel, T. A., & Risse, T. (2010). Governance without a state: Can it work? *Regulation & Governance*, *4*(2), 113-134.

Bulkeley, H., Andonova, L., Betsill, M. M., Compagnon, D., Hale, T., Hoffmann, M. J., . . . VanDeveer, S. D. (2014). *Transnational climate change governance*. New York: Cambridge University Press.

Carlsson, L. (2000). Non-hierarchical evaluation of policy. *Evaluation*, *6*(2), 201-216.

Chan, S., Falkner, R., Goldberg, M., & van Asselt, H. (2016). Effective and geographically balanced? An output-based assessment of non-state climate actions. *Climate Policy*. Online first. http://dx.doi.org/10.1080/14693062.2016.1248343

Chelimsky, E. (2006). The purposes of evaluation in a democratic society. In I. Shaw, J. Greene & M. Mark (Eds.), *The SAGE handbook of evaluation*, (pp. 33-55). London: SAGE.

Chelimsky, E. (2009). Integrating evaluation units into the political environment of government: The role of evaluation policy. *New Directions for Evaluation, 2009*(123), 51-66.

Conley-Tyler, M. (2005). A fundamental choice: internal or external evaluation? *Evaluation Journal of Australasia*, *4*(1/2), 3.

Cooney, K., Dreesen, T., Lees, E. & Titus, E. (2014). Evaluation frameworks: A panel discussion on roles, approaches and gaps. Retrieved from http://www.iepec.org/?p=6804

Crabbé, A., & Leroy, P. (2008). *The handbook of environmental policy evaluation*. London; Sterling, VA: Earthscan.

De Burca, G., Keohane, R. O., & Sabel, C. (2014). Global experimentalist governance. *British Journal of Political Science, 44*(03), 477-486.

Eckstein, H. (2000). Case study and theory in political science. In R. Gomm, M. Hammersley & P. Foster (Eds.), *Case study method: Key issues, key texts* (pp. 119-164). Thousand Oaks, CA: Sage.

EEA. (2016). *Environment and climate policy evaluation.* Copenhagen: European Environment Agency.

European Commission. (1996). *Evaluation: Concrete steps towards best practice across the commission.* (No. SEC 96/659 final). European Commission.

European Commission. (2007). Responding to strategic needs: Reinforcing the use of evaluation. Retrieved from http://ec.europa.eu/smart-regulation/evaluation/docs/eval_comm_sec_2007_213_en.pdf

European Commission. (2013). Strengthening the foundations of smart regulation - improving evaluation. Retrieved from http://ec.europa.eu/smart-regulation/docs/com_2013_686_en.pdf

Feldman, D. L., & Wilt, C. A. (1996). Evaluating the implementation of state-level global climate change programs. *The Journal of Environment & Development, 5*(1), 46-72.

Fischer, F. (2006). *Evaluating public policy*. Mason: Cengage Learning.

Fransen, T., & Cronin, C. (2013). A critical decade for climate policy: tools and initiatives to track our progress. Available: http://www.wri.org/sites/default/files/pdf/critical_decade_for_climate_policy_tools_and_initiatives_to_track_our_progress.pdf.

Furubo, J. E., Rist, R. C., & Sandahl, R. (2002). *International atlas of evaluation*. Transaction Publishers.

Greenwood, J. (2011). *Interest representation in the European Union* (3rd ed.). New York: Palgrave Macmillan.

Guha-Khasnobis, B., Kanbur, R., & Ostrom, E. (2006). Beyond formality and informality. In B. Guha-Khasnobis, R. Kanbur & E. Ostrom (Eds.), *Linking the formal and informal economy: Concepts and policies* (pp. 1-18). Oxford; New York: Oxford University Press.

Haigh, N. (1996). Climate change policies and politics in the European Community. In: T. O'Riordan and J. Jäger, Eds, *Politics of climate change: a European perspective*. New York: Routledge, 155-185.

Hanberger, A. (2012). Framework for exploring the interplay of governance and evaluation. *Scandinavian Journal of Public Administration*, *16*(3), 9-27.

Haug, C., Rayner, T., Jordan, A., Hildingsson, R., Stripple, J., Monni, S., . . . Berkhout, F. (2010). Navigating the dilemmas of climate policy in europe: Evidence from policy evaluation studies. *Climatic Change*, *101*(3), 427-445.

Hayward, R. J., Kay, J., Lee, A., Page, E. C., Patel, N., Payne, H., ... & Valyraki, A. (2014). Evaluation under contract: Government pressure and the production of policy research. *Public Administration*, *92*(1), 224-239.

Hertting, N., & Vedung, E. (2012). Purposes and criteria in network governance evaluation: How far does standard evaluation vocabulary takes us?. *Evaluation*, *18*(1), 27-46.

Hildén, M. (2011). The evolution of climate policies–the role of learning and evaluations. *Journal of Cleaner Production*, *19*(16), 1798-1811.

Hildén, M. (2014). Evaluation, assessment, and policy innovation: Exploring the links in relation to emissions trading. *Environmental Politics, 23*(5), 839-859.

Hildén, M., Jordan, A., & Rayner, T. (2014). Climate policy innovation: Developing an evaluation perspective. *Environmental Politics, 23*(5), 884-905.

Hooghe, L., & Marks, G. (2010). Types of multi-level governance. In H. Enderlein, S. Wälti & M. Zürn (Eds.), *Handbook on multi-level governance* (pp. 17-31). Cheltenham, UK; Northampton, MA: Edward Elgar.

Huitema, D., Jordan, A., Massey, E., Rayner, T., Van Asselt, H., Haug, C., . . . Stripple, J. (2011). The evaluation of climate policy: Theory and emerging practice in Europe. *Policy Sciences*, *44*(2), 179-198.

Hyvarinen, J. (1999). The European Community's Monitoring Mechanism for CO2 and other Greenhouse Gases; the Kyoto Protocol and other Recent Developments. *Review of European Community & International Environmental Law*, *8(2*), 191-197.

Jacob, S., Speer, S., & Furubo, J. (2015). The institutionalization of evaluation matters: Updating the international atlas of evaluation 10 years later. *Evaluation*, *21*(1), 6-31.

Jordan, A., Huitema, D., Van Asselt, H., Rayner, T., & Berkhout, F. (2010). *Climate change policy in the European Union: Confronting the dilemmas of mitigation and adaptation?*. Cambridge; New York: Cambridge University Press.

Jordan, A., & Schout, A. (2006). *The coordination of the European Union: Exploring the capacities of networked governance*. Oxford; New York: Oxford University Press.

Jordan, A. J., Huitema, D., Hildén, M., van Asselt, H., Rayner, T. J., Schoenefeld, J. J., ... & Boasson, E. L. (2015). Emergence of polycentric climate governance and its future prospects. *Nature Climate Change, 5*, 977–982.

Kerr, A. (2007). Serendipity is not a strategy: The impact of national climate programmes on greenhouse-gas emissions. *Area, 39*(4), 418-430.

Levi-Faur, D. (Ed.). (2012). *The Oxford handbook of governance*. Oxford: Oxford University Press.

Leeuw, F. L., & Donaldson, S. I. (2015). Theory in evaluation: Reducing confusion and encouraging debate. *Evaluation, 21(4*), 467-480.

Liverani, A., & Lundgren, H. E. (2007). Evaluation systems in development aid agencies: an analysis of DAC Peer Reviews 1996—2004. *Evaluation*, *13*(2), 241-256.

Löwenbein, O. (2008). The evaluation market in Germany. *Journal of MultiDisciplinary Evaluation, 5*(10), 78-88.

Majone, G. (1989). *Evidence, argument, and persuasion in the policy process*. New Haven, CT: Yale University Press.

Martens, M. (2010). Voice or loyalty? the evolution of the European Environment Agency (EEA). *JCMS: Journal of Common Market Studies, 48*(4), 881-901.

Mastenbroek, E., van Voorst, S., & Meuwese, A. (2015). Closing the regulatory cycle? A meta evaluation of ex-post legislative evaluations by the European Commission. *Journal of European Public Policy*, 1-20.

McDavid, J. C., & Huse, I. (2015). How does accreditation fit into the picture? *New Directions for Evaluation, 2015*(145), 53-69.

Mela, H., & Hildén, M. (2012). Evaluation of climate policies and measures in EU member states–examples and experiences from four sectors. Retrieved from https://helda.helsinki.fi/bitstream/handle/10138/38749/FE19_2012.pdf?sequence=1

Mickwitz, P. (2003). A framework for evaluating environmental policy instruments context and key concepts. *Evaluation, 9*(4), 415-436.

Mickwitz, P. (2006). *Environmental policy evaluation: Concepts and practice*. Suomen Tiedeseura.

Mickwitz, P. (2013). Policy evaluation. In A. Jordan, & C. Adelle (Eds.), *Environmental policy in the EU: Actors, institutions and processes* (pp. 267-286). London; New York: Routledge.

Neslen, A. (2011). EU faces down tar sands industry. Retrieved from http://www.euractiv.com/climate-environment/eu-faces-tar-sands-industry-news-508140

Nickerson, R. S. (1998). Confirmation bias: A ubiquitous phenomenon in many guises. *Review of general psychology*, *2*(2), 175.

Öko-Institut, Cambridge Economics, AMEC, Harmelink Consulting, & TNO. (2012). *Ex-post quantification of the effects and costs of policies and measures.* (No. CLIMA.A.3/SER/2010/0005). Berlin: Öko-Institut.

Ostrom, E. (1990). *Governing the commons: The evolution of institutions for collective action.* Cambridge: Cambridge University Press.

Ostrom, E. (1999). Coping with tragedies of the commons. *Annual Review of Political Science, 2*(1), 493-535.

Ostrom, E. (2005). *Understanding institutional diversity*. Princeton, NJ: Princeton University Press.

Ostrom, E. (2010). Polycentric systems for coping with collective action and global environmental change. *Global Environmental Change, 20*(4), 550-557.

Ostrom, E. (2014). A polycentric approach for coping with climate change. *Annals of Economics and Finance, 15*(1), 71-108.

Ostrom, V. (1999). Polycentricity (part 1). In M. D. McGinnis (Ed.), *Polycentricity and local public economies: Readings from the workshop in political theory and policy analysis* (pp. 52-74). Ann Arbor: University of Michigan Press.

Ostrom, E., Janssen, M. A., & Anderies, J. M. (2007). Going beyond panaceas. *Proceedings of the National Academy of Sciences of the United States of America, 104*(39), 15176-15178. doi:0701886104 [pii]

Pawson, R., & Tilley, N. (1997). *Realistic evaluation*. Thousand Oaks: Sage.

Patton, Q.P. (1997). *Utilization-focused evaluation: The new century text* (3rd Ed.). Thousand Oaks: Sage.

Peters, B. G. (1998). Managing horizontal government: The politics of co-ordination. *Public Administration, 76*(2), 295-311.

Pleger, L., & Sager, F. (2016). Betterment, undermining, support and distortion: A heuristic model for the analysis of pressure on evaluators. *Evaluation and Program Planning*. Available at: http://dx.doi.org/10.1016/j.evalprogplan.2016.09.002

Ratzesberger, P. (2014). Autarkes Dorf Feldheim: Mit eigener Energie. Retrieved from http://www.sueddeutsche.de/wirtschaft/autarkes-dorf-feldheim-mit-eigener-energie-1.2017641

Schoenefeld, J. J., Hildén, M., & Jordan, A. J. (2016). The challenges of monitoring national climate policy: learning lessons from the EU. *Climate Policy*. Online first. http://dx.doi.org/10.1080/14693062.2016.1248887

Somanathan, E., Sterner, T., Sugiyama, T., Chimanikire, D., Dubash, N. K., Essandoh-Yeddu, J., . . . Zylicz, T. (2014). National and sub-national policies and institutions. In O. Edenhofer, R. Pichs-Madruga, Y. Sokona, E. Farahani, S. Kadner, K. Seyboth, . . . J. C. Minx (Eds.), *Climate change 2014: Mitigation of climate change. contribution of working group III to the fifth assessment report of the intergovernmental panel on climate change* (pp. 1141-1205). Cambridge, UK; New York, USA: Cambridge University Press.

Stame, N. (2003). Evaluation and the policy context: The European experience. *Evaluation Journal of Australasia, 3*(2), 37-43.

Stame, N. (2004). The Sixth European Evaluation Society Conference on ‘Governance, Democracy and Evaluation’. *Evaluation, 10(4*), 503-507.

Stame, N. (2006). Governance, democracy and evaluation. *Evaluation, 12*(1), 7-16.

Stame, N. (2008). The European project, federalism and evaluation. *Evaluation*, *14*(2), 117-140.

Stern, E. (2009). Evaluation policy in the European Union and its institutions. *New Directions for Evaluation, 2009*(123), 67-85.

Thompson, T. M., Rausch, S., Saari, R. K., & Selin, N. E. (2014). A systems approach to evaluating the air quality co-benefits of US carbon policies. *Nature Climate Change, 4*, 917-923. Retrieved from http://dx.doi.org/10.1038/nclimate2342

Tosun, J., Koos, S., & Shore, J. (2016). Co-governing common goods: Interaction patterns of private and public actors. *Policy and Society*, *35*(1), 1-12.

Tosun, J., & Schoenefeld, J. J. (2017). Collective climate action and networked climate governance. *Wiley Interdisciplinary Reviews: Climate Change*. Online First. Available at: http://onlinelibrary.wiley.com/doi/10.1002/wcc.440/full

Toulemonde, J. (2000). Evaluation culture (s) in Europe: differences and convergence between national practices. *Vierteljahrshefte zur Wirtschaftsforschung*, *69*(3), 350-357.

Trochim, W. M. (2009). Evaluation policy and evaluation practice. *New Directions for Evaluation*, *2009*(123), 13-32.

Uitto, J. I. (2016). Evaluating the environment as a global public good. *Evaluation*, *22*(1), 108-115.

UNFCCC. (2015). Adoption of the Paris Agreement. Available: http://unfccc.int/resource/docs/2015/cop21/eng/l09r01.pdf

Vedung, E. (2013). Six models of evaluation. In E. Araral, S. Fritzen, M. Howlett, M. Ramesh & X. Wu (Eds.), *Routledge handbook of public policy* (pp. 387-400). London; New York: Routledge.

Vedung, E. (1997). *Public policy and program evaluation*. New Bruswick, N.J.: Transaction Publishers.

Versluis, E., van Keulen, M., & Stephenson, P. (2011). *Analyzing the European Union policy process*. Basingstoke: Palgrave Macmillan.

Voß, J., & Simons, A. (2014). Instrument constituencies and the supply side of policy innovation: The social life of emissions trading. *Environmental Politics*, *23*(5), 735-754.

Waterton, C., & Wynne, B. (2004). Knowledge and political order in the European Environment Agency. *States of knowledge: The co-production of science and social order*, 87-108.

Weiss, C. H. (1993). Where politics and evaluation research meet. *Evaluation Practice, 14*(1), 93-106.

Widerberg, O., & Stripple, J. (2016). The expanding field of cooperative initiatives for decarbonization: a review of five databases. *Wiley Interdisciplinary Reviews: Climate Change.* DOI: 10.1002/wcc.396

Woerlen, C. (2013). Guidelines to climate mitigation evaluations. Retrieved from https://www.climate-eval.org/sites/default/files/studies/Climate-Eval%20Guidelines%20to%20Climate%20Mitigation%20Evaluations.pdf

Tables and Figures

Table 1: Theoretical strengths and weaknesses of formal and informal evaluation

Formal (state-led) evaluation	Informal (society-led) evaluation
Strengths: • better uptake? • inside knowledge, more realistic Weaknesses: • lack of independence, less critical • little publication if evaluations are negative • governments trying to influence evaluators	Strengths: • more critical? • more publication/public discussion? • Lower conflict of interest/influence of governmental actors • Greater scope to be 'reflexive' and focus on side-effects Weaknesses: • Lack of internal knowledge/realism? • Lower uptake/fewer incentives to use results unless there is public pressure • Limited or lopsided funding; not self-organising? • Interest-driven?

Based on: Hildén and colleagues (2014), Huitema and colleagues (2011), and Weiss (1993).

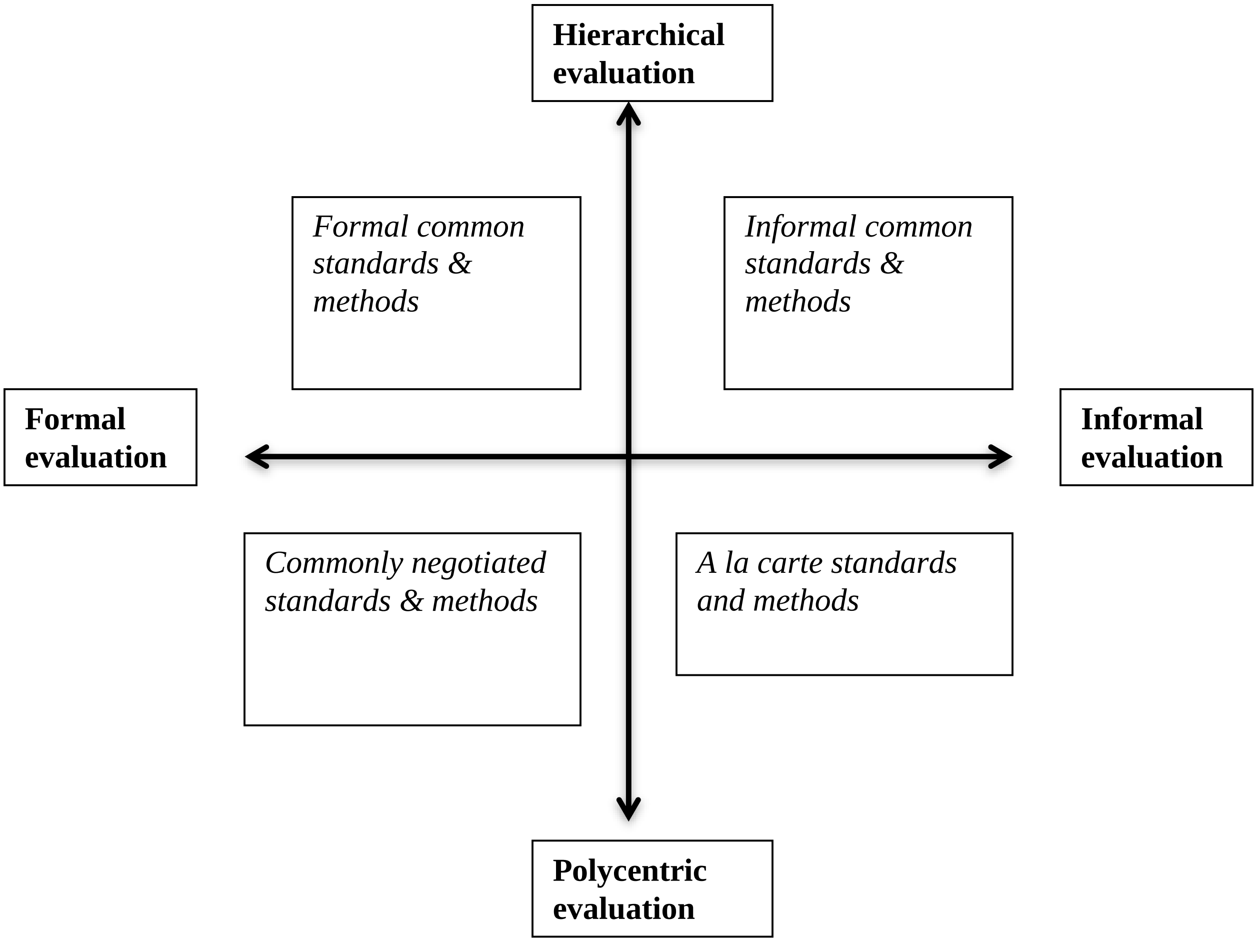
Table 2: Strengths and weaknesses of hierarchical and polycentric evaluation

Hierarchical evaluation	Polycentric evaluation
Strengths: • common standards/comparability • coordination – little duplication, perhaps better knowledge diffusion? • funding/support?	Strengths: • Acknowledges self-organising capacities; more localised 'ownership' of evaluation activities

• More evaluation capacity Weaknesses: • insensitivity to context → key in policy success • may be perceived as 'policing' → actors unwilling to evaluate/cooperate? • High risk of systemic failure; difficult to address problems/change course • What if there is no central actor with sufficient resources to evaluate (e.g., UNFCCC?)	• 'Uncomfortable' evaluation results more likely to emerge, particularly if addressing influences from different governance levels • sensitivity to context • low risk of systemic failure → if one part of the system fails, there are still many others • Does not rely on a single actor Weaknesses: • Multiple standards, hard to compare? • Collective action problem – policy evaluation/knowledge as a common pool resource? • Lack of resources at lower levels; lack of evaluation capacity

Based on Elinor Ostrom (2010; 2014) and Vincent Ostrom (1999).

Figure 1: Approaches to governing policy evaluation


Hierarchical evaluation
Formal common standards & methods
Informal common standards & methods
Formal evaluation
Informal evaluation
Commonly negotiated standards & methods
A la carte standards and methods
Polycentric evaluation