Journal of MultiDisciplinary Evaluation
Volume 21, Issue 50, 2025

ISSN 1556-8180
http://www.jmde.com

Building a Mountain of Evaluative Evidence, 2004–2014

Rob D. van den Berg
Honorary Associate, Institute of Development Studies

Background: The book *From Studies to Streams* was for me an eye-opener when I worked as director of the Independent Evaluation Office of the Global Environment Facility (GEF). Right from the start in that position I was working to gather as much evaluative evidence as possible, mix this with knowledge and insight, and look at how to deliver recommendations and insights to the GEF. *From Studies to Streams* inspired me to work toward a potential mountain of evidence to inform and inspire the replenishment meetings of the GEF. While the book provided an analogy of evaluation insights and evidence streaming down to the ocean, I more felt that the knowledge gathered in the Overall Performance Studies would be moving up to reach higher levels of decision-making. At the top of the mountain of evidence and insight, the GEF replenishment meetings would decide on the goals of the fund in the next 4 years.

Purpose: This paper aims to provide a historically accurate account of how evaluations at different levels of GEF-funded activities were used to inform higher-level evaluations, leading to an integrative perspective of achievements. Evaluations incorporating both scientific and national/local perspectives aimed to capture findings and insights at all levels of the GEF and its partners. While this was not the stream downward to the ocean, it could be likened to steam that gradually wafted up to the pinnacle of GEF decision-making.

Setting: The partnership of the GEF with multilateral banks, five UN organizations, and the 120 recipient countries of GEF funding. While all agencies and countries had their own evaluation policies, agreement was reached about a minimum number of common elements that would be reported on. All evaluations that touched upon GEF issues were studied to extract insights relevant for the GEF. This was combined with knowledge generated through the GEF Scientific and Technical Advisory Panel (STAP), and any relevant knowledge available through literature and expertise.

Intervention: Not applicable.

Research Design: Not applicable.

Data Collection and Analysis: Not applicable.

Findings: It turned out to be possible to create a flow of evaluative evidence and other insights and knowledge regarding interventions from the many varieties of evaluation that were undertaken with the GEF and its many partners. This led to a veritable mountain of evidence that was presented every 4 years to the replenishment meetings for the GEF. While the GEF is a relatively unique international funding organization, it turned out to be possible to make use of the evidence generated at various levels of the partnership and by different actors, along the lines of *From Studies to Streams*.

Keywords: *Global Environment Facility, studies, comprehensive evaluation, integrated approaches, evidence streams, evaluation influence, evaluative evidence, knowledge.*

The article by Nicoletta Stame (see this volume) reveals the thinking that led to a proposed move from individual studies to streams of evaluative evidence. While the book on this issue was published in 2006, I had already started thinking about this when I worked at the Dutch Ministry of Foreign Affairs as director of evaluation from 1999 to 2004. In that position I was fully convinced that the results of one study would not be sufficient to lead to any changes that were advocated in that specific evaluation report. Apart from knee-jerk reactions along the lines of "Oh, we stopped doing that ages ago," or "We knew that all along but have to live with that political reality," I noticed that evaluations succeeded in convincing management, ministers, Parliament, and the public *if* they consistently found over time that some practice could and should be improved. For example, the demise of old-fashioned technical assistance around the 1990s was caused by a series of evaluations of technical assistance that one after the other pointed to the caricature of Western old men telling local people how their problems should be solved, often without any success, as the Western recipes could not be applied in Southern circumstances and cultures. [1] These evaluations demonstrated that they could contribute to a better understanding of what went wrong, through similar and reproducible findings. I felt that on other topics there was also an emerging consensus on "good" versus "bad" practices, but this did not (yet) lead me to work on "streaming" evidence in Dutch evaluation practice.

When I was appointed director of the newly independent office for evaluation of the Global Environment Facility in September 2004, I was confronted with three factors that made it much easier to think in terms of continuity rather than discontinuity. The first one was that the GEF was in a unique position as a financial mechanism of the multilateral environmental agreements. [2] These conventions led to negotiations to achieve clarity on what should be funded. Activities on the environment/development nexus were thus rooted in international law; countries had put their signatures on these agreements. This is not the case in almost anything else in international cooperation. Poverty alleviation, for example, is a question of good will rather than of an international agreement signed by almost all UN member states. The conventions, on the other hand, provide a linkage to what a financial instrument such as the GEF was supposed to fund, with what governments were supposed to prioritize. This spread over more than 120 countries that were eligible for GEF funding.

The second factor was a certain level of standardization of evaluation practice in multilateral institutions, which makes it easier for evaluative findings to be shared and recognized beyond the organization that produces them. Three international forums were working on establishing best international practices for evaluation: the evaluation network of the OECD/DAC, mainly representing bilateral evaluation agencies; the evaluation cooperation group of the multilateral financial institutions; and the UN evaluation group, which in these years worked on norms and standards for evaluation in the UN. In my time, the GEF operated through five banks and five UN organizations, thus working with the standards of two of the three groups.

On top of that, the nexus between development and environment in which the GEF operated, was a subject of methodological development. While the social sciences were often used in development evaluation, the environmental side needed to be covered with methodologies most often originating in the Earth sciences. Multidisciplinary evaluation work brings its challenges! The Scientific and Technical Advisory Panel (STAP) supported the council and secretariat of the GEF in adopting methodologies to fulfill its mandate. The demand for independent evaluation grew in the GEF in the early 2000s, and I was appointed as the first director of an evaluation office that could set up a system to evaluate the GEF network. Every 4 years, the GEF has been the subject to Overall Performance Studies to inform each replenishment process. This became the focus of our evaluation programming.

All in all, these three factors would perhaps have been insufficient if the system in place to support interventions and policies had not strongly supported sharing knowledge, insight, and evaluative evidence at different levels. To make that

[1] For a good overview of the history of technical assistance, see Cox & Norrington-Davies (2019).

[2] Currently focused on the Convention on Biological Diversity, the UN Framework Convention on Climate Change, the Stockholm Convention on Persistent Organic Pollutants, the UN Convention to Combat Desertification, the Minamata Convention on Mercury, and the Biodiversity Beyond National Jurisdiction Agreement. Furthermore, the GEF provides support to the Montreal Protocol on Substances that Deplete the Ozone Layer and multilateral agreements on international waters and transboundary water systems. More information about the multilateral environmental agreements can be found at https://oneplanetnetwork.org/SDG-12/multilateral-environmental-agreements.

possible, the GEF had a consistent portfolio, had agreement on how the achievements of its investments were to be monitored and reported, and had a process in place for evaluative streams to become useful (as the replenishment processes demonstrates in the GEF).

But before going into this modus operandi, I need to point out two important differences with the position taken in *From Studies to Streams*. As Nicoletta describes it, the book that she and Ray Rist edited came out against individual studies and focused on streams to overcome the irrelevance of individual studies. The first issue is that individual studies sometimes actually do play a key role. Let me give an example of this. In the years before I came to the GEF, two different expert opinions dominated the global support that could be given to protected areas for conservation purposes. The stream of thought that the first group adhered to was that conservation of nature should take shape by keeping people away from protected areas. For them, allowing people to move into these areas would mean that these would fail, as humans would kill animals, turn primary forests into farmlands, and so on. The only safe environment was seen as an environment without humanity. The global goal (at the time) was to create protectorates on 10% of the Earth's surface. From these pockets of nature humanity had to be banned, as they firmly believed science and expertise had shown and demonstrated sufficiently.

A second stream of thought and of different scientific findings was less dogmatic about humanity and proposed that where humanity had a tradition to manage nature, it would create global benefits, such as increased or safe biodiversity and sustainable production of plants and animals in safe surroundings. The GEF Council aimed to end this discussion by commissioning a study, to be undertaken by evaluators, that would look at whether *local benefits* (for the people) would undermine, or be neutral, or would increase *global benefits* (for the environment). If the study showed that local and global benefits could both be achieved, they would conclude that people would be welcome (within reason) in conservation areas.

This study attracted many donations from bilateral donors who wanted the fight between the two groups of scientists/conservationists to be solved. More than $4 million was raised. The study took several years to complete and led to a report in 2006 (GEF EO, 2006). It concluded that for many areas of GEF support, local and environmental benefits are interlinked. You cannot have one without the other. The study laid to rest a controversial point in GEF practice. It was one study with a huge impact on future practice. I have more examples of where individual studies can make a difference, but this is just one illustration that not all individual studies should be abandoned.

The second issue with the concept from studies to streams is that streams flow down from a source, sometimes from a mountain, and usually end up in the ocean. From small to big, from many different points (all part of the catchment area of a river) to a mighty stream that through a delta reaches the ocean. Yet this is not what is happening in the case of bringing evaluative knowledge and evidence to the GEF network. Knowledge is gradually built up; it is not flowing down. Our analogy was more of a mountain, where the knowledge would gradually be brought to the top of the mountain by sherpas who carefully listened to the knowledge from down below, gathering insights that would make sense at the top of the mountain. And to take the analogy of the mountain one step further, more than flatland (like the Netherlands), mountains are synonymous with great diversity of vegetation (or lack of it) and landscape. The mountain ranges in Costa Rica are an extreme example of a spectrum of ecosystems from bottom to top, from tropical to temperate rain forests to the craters of volcanoes. The road to the top comes with many discoveries. To conclude: The analogy of a stream does not seem appropriate for GEF evaluative knowledge. It should be more an analogy of steam: hot air generated in the tropical forests at the foot of the mountain range, wafting up as steam to higher areas, where most of this either disappears or returns as rain or mist. We thus arrive at language that addresses knowledge gathering at the GEF both in terms of a stream and in terms of steam, thus explaining why the stream goes up instead of down.

The evaluation practice of the GEF network can best be described as a uniform approach of diversity. Uniform in the sense that reporting of evaluative evidence would be according to rigorous standards, and diversity in the sense that each evaluation would focus on the relevant evidence for the specific situations and interventions evaluated. The uniformity enabled us to gradually work toward steam rising up of evaluative evidence. The diversity ensured that the findings would reflect local circumstances and unique implementation mechanisms. The nature of the synthesis at higher levels is thus not a strict synthesis study according to Campbell or realist rules, but a combination of meta evaluation and meta-analysis of "knowledge" and "information," as they were emerging in the stream of steam. The famous Table 1.1, "Types of Streams of Evaluative Knowledge," in the book *From Studies to Streams* (Rist & Stame, 2006, p. x) applies to the GEF for both mixed types of information and evaluative knowledge, as noted in Cells 2 and 4.

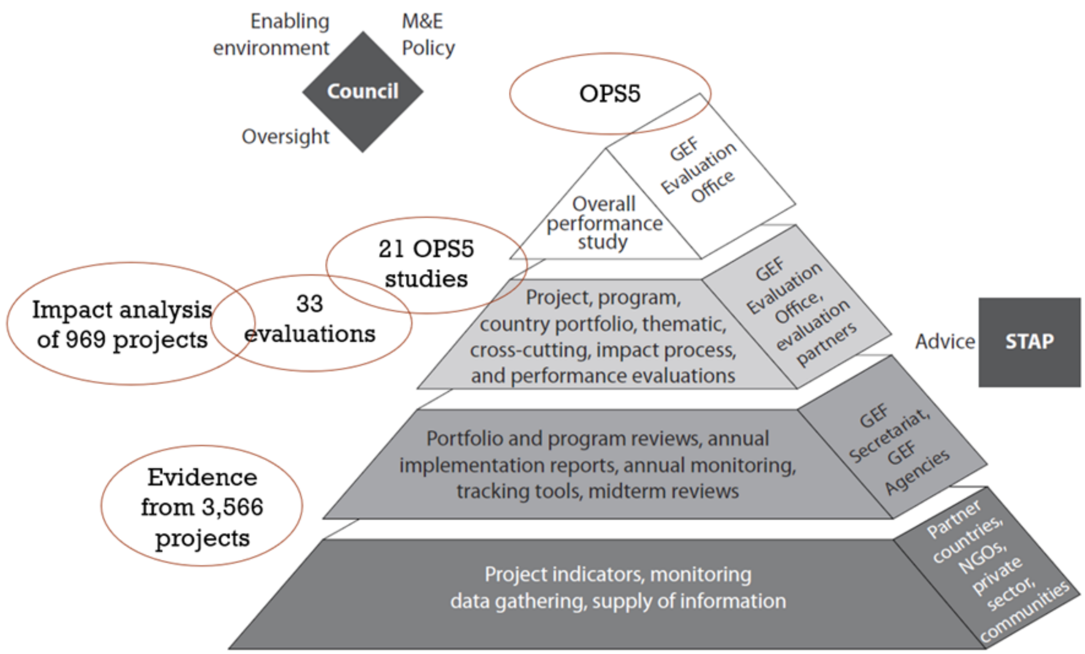
While all the implementing and executing agencies, countries, and sometimes local organizations had their own evaluation policies, we were able to reach agreement on minimum standards of reporting that would enable us to add up and compare the findings of all evaluations of activities funded by the GEF. This was the first (humble) step toward a stream of evidence, or to put it in words more in line with the GEF analogy of a mountain, a stream of uplifting steam.

Furthermore, many agencies had their own thematic and country portfolio–related evaluations, in which their GEF portfolio was included. On top of that, the GEF independent evaluation office would also initiate sectoral and thematic as well as country portfolio evaluations, which reported on how agencies performed and whether GEF policies were successful or could be further improved. Impact was a special issue. While some randomized controlled trials took place, most impact of the GEF, especially on the environment, could not be measured in that way. Instead of impact as defined by what works and what doesn't, impact in the GEF was defined as the final state that was aimed for—for example, an ecosystem restored to its former glory. For this, other measurements and scientific research were needed. Since final goals were long-term efforts, impact was reported as "progress towards impact."

The aim in all these evaluations was to understand what was happening at the many different levels of action and for the many different stakeholders involved in these actions. All evaluations were focused on the specific audiences identified and were discussed with them. We could say that evaluations had common elements, as discussed above, but focused at the same time on the specifics of the situations and interventions to be evaluated, and provided feedback at the appropriate levels and to the relevant stakeholders and beneficiaries.

The common elements enabled us to report to the replenishment of the GEF on the overall achievements in the GEF partnership, but where specific evaluations generated knowledge that provided new insights, this was reported as well. We aimed to integrate all findings in the 4-year cycle of Overall Performance Studies. Figure 1 shows the mountain of evidence that we managed to include in the last study that I was responsible for, OPS5. It shows the different layers of evaluative evidence. At the base is the monitoring reporting system, focusing on indicators. The data from this would be used with evidence from 3,566 GEF projects, all originating from the GEF agencies and the GEF secretariat through project terminal evaluations and annual implementation reviews. A separate analysis of progress toward impact was available for 969 projects of the GEF. Evaluation partners in the GEF and the independent evaluation office itself added 33 evaluations to the next level of the pyramid, and 21 specific studies of OPS5 were initiated to bring evidence together throughout the pyramid, leading to the overall study publication (GEF IEO, 2014), which was presented to the replenishment process in 2014, at the end of my period at the helm of the Independent Evaluation Office.

Figure 1. The Stream of Evaluative Steam Moving Up to OPS5



Note. Figure prepared by Rob D. van den Berg for Kings College London, 2016.

An important point to make is the greater uniformity of action in the GEF, with a limited number of modalities and international agreement on goals and aims. But also, the scientific support as achieved through the STAP is important. For example, STAP advised the GEF on how to calculate CO_2 emission reductions, so important for climate change. This led to uniformity in data gathering and reporting that was quite extraordinary. As my co-writer and I found previously (Van den Berg & Cando-Noordhuizen, 2017), in 2015 several major international evaluations were undertaken to show what had been done so far on reducing CO_2 emissions, but only two organizations were able to present their track record in doing so: the GEF and Norway's International Climate and Forest Initiative (NICFI). Why only two, and why not the World Bank, the Climate Investment Funds, the Interamerican Development Bank, the Asian Development Bank, and the Swiss International Cooperation in Climate Change? All those had overall evaluations of their efforts, yet these were not able to report on the overall achievements, because they had no agreed-upon calculation process that was applied by all their projects. A special place should be accorded to the UN-REDD+ program, that was evaluated at the same time. It had a uniform portfolio and even agreed upon ways of calculating CO_2 emission reductions, but its evaluation did not look into what the overall achievement of the UN-REDD+ program amounted to—a lost opportunity.

Reaching agreement on the metrics of evaluation reports enables comparisons and portfolio understanding, but this does not guarantee that the evaluative evidence thus identified is automatically accepted. OPS5, the first of the synthesis studies of the GEF, became a victim of a disconnect between what evaluative evidence was saying and what the CEO aimed for. A new CEO had come to the GEF, and she aimed to tell the replenishment that new thematic programs were needed, as the traditional GEF activities, according to her, had no impact. Furthermore, she told the donors that the GEF was an efficient organization. OPS5 provided evaluative evidence that this vision was not correct. According to our evaluative studies, the GEF was in danger of becoming inefficient, unless appropriate action would be taken. Furthermore, our evaluative evidence showed that impact expectations of the GEF-funded interventions were positive and well embedded in national priorities. This difference in vision was not resolved at the time—or was perhaps satisfactorily negotiated out of the way by the replenishment not taking over the full recommendations of the CEO, while also not taking over all evaluative recommendations of OPS5.

Whether evaluations, monitoring, and other information sources can deliver streams of evidence that make sense at higher levels is thus also dependent on whether the organization, fund, or program to be evaluated has a consistent portfolio, has agreement on how its achievements are to be monitored and reported, and has a process in place for evaluative streams to become useful (as the replenishment processes demonstrated in the GEF). While the first effort to bring a stream of evaluative evidence to the replenishment was not yet fully successful, the next section of this article will tackle how the evaluation office under Juha I. Uitto continued to build up this stream of evidence and how it influenced GEF practices.

References

Cox, M., & Norrington-Davies, G. (2019). *Technical assistance: New thinking on an old problem*. Open Society Foundations, Agulhas, Applied Knowledge. https://tinyurl.com/ycysb2za

GEF EO (2006). *The role of local benefits in global environmental programs*. Global Environment Facility Evaluation Office. https://tinyurl.com/4snv4rw3

GEF IEO (2014). *Fifth overall performance study of the GEF. Final report: At the crossroads for higher impact*. Global Environment Facility Independent Evaluation Office. https://www.gefieo.org/evaluations/ops5-final-report.

Rist, R., & Stame, N. (Eds.). (2006). *From studies to streams: Managing evaluative systems*. Transaction Publishers.

Van den Berg, R. D., & Cando-Noordhuizen, L. (2017). Action on climate change: What does it mean and where does it lead to? In J. I. Uitto, J. Puri, & R. D. van den Berg (Eds.), *Evaluating climate change action for sustainable development* (pp. 13–33). Springer Open.