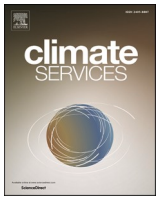
Climate Services 38 (2025) 100549

Contents lists available at ScienceDirect

Climate Services

journal homepage: www.elsevier.com/locate/cliser



Developing and testing an evaluation framework for climate services for adaptation

Eva Boon [a,b,*], Nellie Sofie Body [c], Robbert Biesbroek [d]

[a] Wageningen University & Research, Earth Systems and Global Change Group, Droevendaalsesteeg 3 6708 PB Wageningen, the Netherlands
[b] Foundation Climate Adaptation Services, Bussummergrindweg 1 1406 NZ Bussum, the Netherlands
[c] Norwegian Geotechnical Institute, Oslo, Norway
[d] Wageningen University & Research, Public Administration and Policy Group, Hollandseweg 1 6706 KN Wageningen, the Netherlands

HIGHLIGHTS

- The framework evaluates climate services using agreed-upon success criteria.
- This framework supports climate service development, research, and evaluation.
- Good evaluation is done best when integrated in service development.
- The framework requires clearly defined users and goals for robust assessment.

ARTICLE INFO

Keywords:
Climate services
Climate change adaptation
Evaluation framework

ABSTRACT

Climate services are increasingly developed and used to plan for climate change adaptation, but their success is poorly evaluated. A main reason is that an operational framework to support climate service researchers and practitioners pursuing evaluation is lacking. This study addresses this gap by developing and testing a robust and systematic evaluation framework in three steps. First, we designed a framework by operationalising agreed upon criteria for assessing climate service success. Second, the framework was tested in two climate service cases. Third, the usability, credibility, and transparency of the framework was assessed by climate service researchers and practitioners, including those engaged in the cases.

Our findings show that developed framework offers a standardized approach to evaluation, providing indicators, metrics, and guidance that enable the evaluator to provide a quantitative rating for each criterion. However, the robustness of ratings in the two cases was compromised due to limited interaction with targeted users during the development process and lack a of clear goals set from the beginning. This hampered incorporating the perception of a representative group of users and measuring impacts. Overall, the framework was considered usable by researchers and practitioners for various applications, including using it as design criteria, to facilitate learning, to guide development, and to support monitoring and evaluation. While generally perceived as credible and transparent, the framework would benefit from further testing and elaboration into practical materials. The study highlights that evaluation is done best when evaluation criteria are considered early in the development of the climate service.

Practical implications: Climate services are seen as important means to support and accelerate adaptation action. While investments in climate service development and use are increasing, their evaluation typically falls short. One reason for this is the lack of a sound evaluation framework. This study aimed to develop a robust and systematic evaluation framework that can be used in both science and practice settings. The framework was tested in two implemented climate service cases, and evaluated by climate service users, practitioners, and researchers, as well as by the evaluators themselves. Supplementary file 2 provides the framework, and an accompanying protocol describing important process steps to apply it. It also offers guidance on how to consider the success criteria during the development stages of a climate service, through guiding questions and a checklist. Here we present the practical implications of this study by (1) outlining the basic principles of the framework,

* Corresponding author.
E-mail address: eva.boon@wur.nl (E. Boon).

https://doi.org/10.1016/j.cliser.2025.100549
Received 5 September 2024; Received in revised form 3 February 2025; Accepted 3 February 2025
Available online 12 February 2025

summarizing the results of (2) testing and (3) evaluating the framework that have most practical relevance, and (4) highlighting suggestions for improving evaluation practice.

1) Basic principles of the framework are:

- It can be used for different types of evaluation (e.g. summative, formative, developmental) and applied to the broad range of possible climate services.
- It is based on 12 success criteria selected in a Delphi study, where experts evaluated which elements are most relevant to define the success of climate services for adaptation (Boon et al., 2024). If deemed necessary for a specific climate service or context, criteria can be added.
- It offers a total of 20 indicators with supporting metrics and directions to measure the criteria. Indicators were selected based on literature review, considering the most robust approach for measurement while dealing with time and budget restrictions.
- Each criterion is evaluated on a scale from 1: unsuccessful to 5: successful, allowing easy comparison between climate services and monitoring over time. The robustness of the rating is assessed by considering the representativeness of the sample and the extent to which evidence was validated through multiple sources.

2) Testing the framework in two cases shows:

- The framework was usable to evaluate the criteria consistently, supported by clear metrics and instructions for measurement.
- Challenges emerged for evaluating those indicators that require a clear definition of targeted users and goals, and for those that are measured through user perception. In both cases, users and goals were described only in general terms, which made it difficult and sometimes impossible to measure results for these indicators. Furthermore, the robustness of many ratings was compromised due to the difficulty in accessing a representative group of targeted users.
- The evaluation results, including identified learnings, were recognized and appreciated by the involved stakeholders.

3) Evaluating the evaluation framework by climate service users, practitioners, and researchers shows:

- The framework was considered usable for various applications, such as including it in the terms of reference of calls for tenders, developing business models, using it as design criteria, guiding development processes, supporting monitoring and evaluation, and facilitating learning about what works and what doesn't work.
- It was considered credible and transparent, although it needs further testing in different types of services and contexts, and may require further development of easy-to-use evaluation materials.
- Especially the climate service producers and practitioners valued the framework.

4) Suggestions for improving evaluation practice:

- The study highlights once again that good evaluation is done best when it is integrated early in the development process of a climate service. This approach not only allows for efficient data collection, but also helps establish more robust ratings by clearly defining users and goals of the climate service and setting up user interaction channels. This may lead to more successful services.
- To stimulate the uptake of the framework and foster a culture for evaluation we see two promising pathways: 1) promoting the use of the success criteria as a helpful tool to guide and structure the climate service development process. Increasing awareness of the criteria may pave the way for more systematic efforts to evaluate the services; 2) promoting the necessity for evaluation, for example to be able to mitigate misguided or ineffective services. This could be done through mandatory use and evaluation of the success criteria through design or reporting requirement by commissioning parties.

Introduction

Climate services for adaptation are poorly evaluated in both research and practice (Boon et al., 2022; Jahan et al., 2023; Tall et al., 2018). Without robust and systematic evaluation, claims about the success of climate services remain case specific, anecdotal, or unjustified. The consequences are far-reaching and can hinder effectiveness in climate action, erode trust in science and policy, and result in misguided priorities. In a context where the number of scientific publications on the topic is rapidly growing (Boon et al., 2022; Larosa and Mysiak, 2019), and where investment in developing and applying new climate services is increasing (IPCC, 2022), the need for their evaluation is evident. Evaluation enables reporting on successes and failures, researching what type of climate services works, when and why, and can improve climate services before, during, and after their development.

Reasons for limited or poor evaluation are plentiful, including lack of budget and capacity for evaluation in climate service projects, diverse and conflicting views on evaluation criteria, resistance to rigorous evaluation due to fear of criticism or negative findings, underappreciation of the benefits of evaluations, short term focus of many projects and programs, and methodological challenges related to for example data collection and establishing causality (e.g. see Jahan et al., 2023; Tall et al., 2018). Moreover, an operational and standardized evaluation framework is lacking (Bremer et al., 2021).

Some climate service evaluation frameworks exist, but many focus on a specific part of the service, for example the co-creation processes in climate service development (Schuck-Zöller et al., 2022; Visman et al., 2022; Wall et al., 2017) or the quality of the knowledge in the service

(André et al., 2021). A notable example is by Vaughan and Dessai (2014) who offer design elements for a comprehensive evaluation framework, which is further developed into evaluation metrics by Jahan et al (2023). These elements, however, are more focused on what contributes to success (enabling conditions) rather than defining the outcome. Moreover, the metrics are designed for quick and broad evaluation using data that is readily available, not including experiences and perceptions of the users and producers. Another comprehensive framework is offered by Bremer et al (2021), who developed a checklist for stakeholders to co-create quality criteria to assess a climate service, covering input, process, output, and use. This framework supports developing a context specific meaning of success and evaluating success accordingly but is less useful for systematic comparison of the success of different climate services. In short, there is a need for an operational evaluation framework that is sufficiently detailed to provide insight on the individual case level, and general enough to apply it to the broad range of climate services enabling learning across cases. Moreover, the framework needs to be usable to support the much-needed evaluation culture and practice.

This study therefore aims to develop and test a framework that allows robust and systematic evaluation of climate services that can be practically applied in both project and research settings. Robust here means that it is conceptually coherent and that the best available methods are used to measure success criteria. Systematic refers to that it guides the evaluation of different services in the same comprehensive way, allowing comparison between climate services as well as monitoring over time. Practically applicable is about keeping it simple and pragmatic. Given the range of possible climate services, we aim to develop a framework that is flexible and can be used for different types of evaluation (e.g. formative, developmental, summative).

Earlier work offers key anchor points for developing and testing this framework. Boon et al (2024) use a Delphi study to identify 12 criteria that experts agree are key to defining the success of climate services for adaptation, see Table 1. The criteria relate to three conceptual categories: the production process (P, 1 criterion), characteristics and qualities of the climate service itself (C, 6 criteria), and results that follow from using or producing the climate service (R, 5 criteria). The definition assumes a broad understanding of climate services, referring to the development, delivery, and/or use of climate-related knowledge products and/or processes in context of climate change adaptation. It focuses specifically on those services that aim to support long-term planning and investment decisions, rather than strategies addressing forecasted climate events within a decade. Examples of climate services for adaptation include, climate change impact tools, climate stories, adaptation guidance documents, assessment of adaptation action effectiveness, climate projections, and serious games focused on climate impacts and adaptation action (Boon et al., 2024; Findlater et al., 2021; Street, 2016; Weichselgartner and Arheimer, 2019).

This paper is structured as follows. The next section details the methods used for developing, testing, and evaluating the framework. The three subsequent sections present the evaluation framework (section 3), the results of testing it in two climate service cases (section 4), and an evaluation of the framework itself (section 5). We conclude the paper with a discussion and suggestions for future research.

Table 1
Success criteria of climate services for adaptation.

P1. Interaction between users and producers is tailored to context
C1. Timely delivery
C2. Accessible climate service
C3. Credible information
C4. Relevant information
C5. Acknowledgement and communication of uncertainty
C6. Communication format is tailored to users
R1. Climate service increases users' understanding of an issue
R2. Users build the capacity for using services
R3. The climate service has tangible or intangible benefits for the user
R4. Establishment of trust between users and producers
R5. Better decision-making for adaptation

Methods

Research design

The study followed three consecutive steps in which the evaluation framework was designed, tested, and evaluated (see Fig. 1). First, a framework was designed to evaluate the success of climate services, using the 12 climate service success criteria as defined by experts (Boon et al., 2024). Second, the framework was tested in two climate service cases: climate stories for the municipality of Milan, Italy and the municipality of Lillestrøm, Norway. Finally, based on the test in the two cases, the evaluation framework was evaluated using three criteria: usability, credibility, and transparency (United Nations Evaluation Group, 2017). The goal of this step was to explore the added value of the framework for both research and practice, and to identify potential improvements of the framework. The next sections further detail the methods used in each step.

Designing an evaluation framework for climate services

First, scientific literature was reviewed to identify best practices for operationalizing the 12 success criteria. This review was not limited to the field of climate services, drawing in cases from literature where evaluation is more developed, such as the field of transdisciplinary research or science-practice projects (e.g. see Belcher et al., 2016; Walter et al., 2007). The most suitable indicator or combination of indicators for each criterion was explored, guided by three key questions:

1. *What* is/are (a) possible indicator(s) to measure this criterion, also considering the broad range of climate services?
2. *How* can the indicators be measured?
3. *Who* can measure this indicator best?

Additionally, to ensure broad applicability of the framework, each indicator was reviewed on the extent to which it would be possible apply it to a service that is highly data intensive (e.g. technical tool with climate variables) and a service that is process intensive (e.g. a workshop method). For each selected indicator, a 5-point scale was developed, with guidelines on how to choose a rating. Score 1 corresponds with an unsuccessful outcome, and 5 with a successful outcome. Using a quantitative scale allows for comparison between different climate services, instead of only understanding the success of a single case.

In addition to the evaluation framework, a list of process steps that were deemed critical or helpful for the evaluation of climate services were defined, including how to apply the operationalized criteria and synthesize the results into an evaluation. These process steps were summarized in the protocol for evaluation. Supplementary file 1, section 1 describes how relevant norms for evaluation were included in the design of the framework and protocol (United Nations Evaluation Group, 2017).

Testing the evaluation framework: Two climate story cases

The climate stories of Milan and Lillestrøm were used as testcases for the evaluation framework. The two cases were selected from six cases in the Horizon Europe research project REACHOUT*, which offered us the opportunity to closely follow the development of the projects' climate services and to evaluate them. The cases are similar in terms of the development process, service type, and targeted users, but differ in their context of development and implementation. In both cases, the climate services developed are climate change adaptation themed stories which the cities use to reach a defined audience with a specific message. Both stories were developed using the 'Hero's Journey', a narrative structure

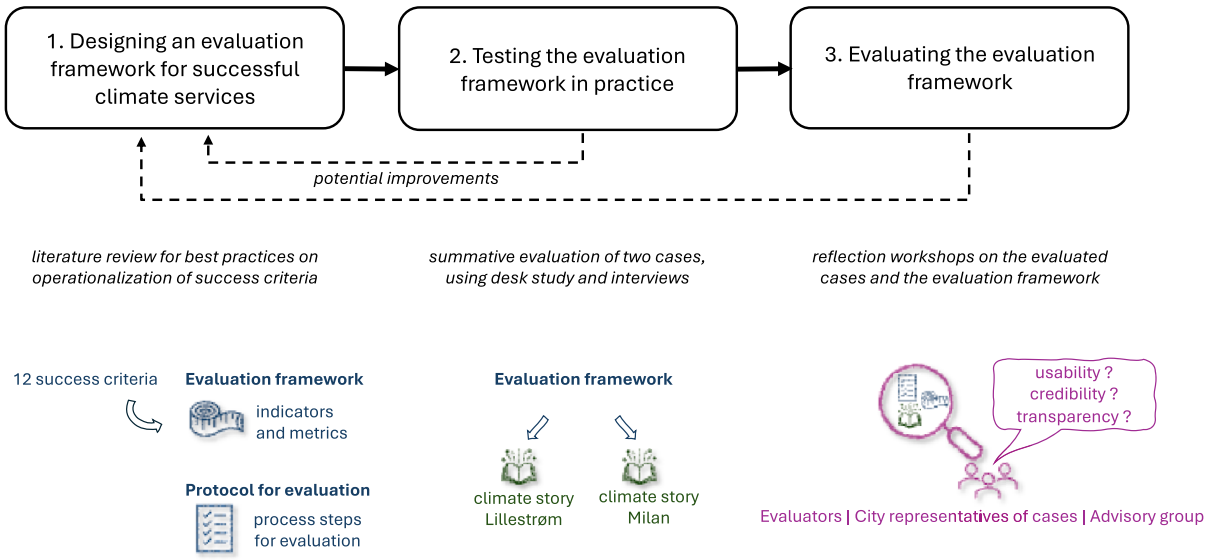


Fig. 1. Research design indicating the three steps in the research.

well known from fairytales (Barel, 2020). The motivation for using storytelling is that by incorporating human and relatable elements into climate (adaptation) knowledge, an emotional and personal connection is formed with the users. This in turn can lead to a greater impact than merely presenting general information. The stories are presented through a scrollable webpage, combining text, graphs, maps, infographics, photos, videos, and illustrations of characters. Milan and Lillestrøm have different climate challenges, city size, and maturity in adaptation planning and the use of climate services. While a pragmatic selection, the cases were considered suitable testcases because it was expected that most, if not all, indicators could be tested. This assessment was based on them having a higher likelihood of sufficient project time to measure potential impacts, as well as their diversity in addressing various dimensions of climate services. These included co-production, integration of both data and softer communication aspects, and all-over comprehensiveness.

Following the protocol for evaluation, we defined the goals of the evaluation, mapped relevant stakeholders, and identified opportunities for data collection. The framework developed in the previous step also allowed for adding new criteria, and the project team decided to add '*Fair Process*' as an evaluation criterion (see also supplementary file 1, section 2). Fair process, here, refers to the engagement of all relevant stakeholders in developing the climate story and integrating their knowledge in an appropriate way.

The evaluation started at the time that the climate stories were just finished (Milan) and to be finished in a few months (Lillestrøm). Data collection and analysis alternated and was done in several rounds of interviews with the key stakeholders and desk study to allow for validating preliminary findings and resolving loose ends, see Table 2 for a timeline. The key stakeholders included two groups: 1) the city representatives and 2) the producers of the story.

The city representatives were each interviewed in two rounds (online), and the producers were questioned through a group interview with the producers of all climate stories (in-person session). All invited stakeholders participated in the evaluation. The semi-structured interviews were informed by the evaluation framework, and included both open questions to collect general experiences and views, and closed questions to gather success perceptions on specific criteria using a 5-point Likert scale. The interview guides and more details about the interviews are included in supplementary file 1, section 3.

The desk study mainly involved analysis of documents describing the process and (intermediate versions of) the climate story itself. This included meeting notes, project deliverables, a document describing first ideas for the story based on a brainstorm between city representatives and producers ('climate story intake template'), versions of the climate story design ('story board'), and an excel document where the producers reported the meetings they had with the city representatives, including the main purpose and outcomes of the meeting. User statistics (web analytics) and a survey integrated in the story were also analyzed.

The transcribed interviews and the results from the desk study were analyzed to build and validate summaries of the production process, the climate story product, and any results of using and producing the service, in line with the conceptual categories. The summaries were used to establish tentative and final ratings for each of the indicators.

Evaluating the evaluation framework

The evaluation of the framework was done by various people: the evaluators themselves, the city representatives involved in the evaluated cases, and the wider group of city representatives, tool developers, researchers, and knowledge brokers engaged in the project ('advisory group'). Evaluation included collecting general views and experiences

Table 2
Timeline of data collection and analysis for the climate stories of Milan and Lillestrøm.

Milan		Lillestrøm	
Spring 2023:	Analyze climate story and documents	Summer 2023:	Analyze climate story and documents
Summer 2023:	Interview city round 1 (n = 3)	Summer 2023:	Interview city round 1 (n = 2)
Winter 2023:	Analyze documents	Winter 2023:	Analyze documents
Spring 2024:	Group interview producers (n = 2)	Spring 2024:	Group interview producers (n = 3)
Spring 2024:	Interview city round 2 (n = 1*)	Spring 2024:	Interview city round 2 (n = 2)
Spring 2024:	Analysis of survey (n = 0) and user statistics	Spring 2024:	Analysis of survey (n = 9) and user statistics
Spring 2024:	Compile evaluation	Spring 2024:	Compile evaluation
Summer 2024:	Presentation and discussion of evaluation results	Summer 2024:	Presentation and discussion of evaluation results

*city representatives from round 1 left the municipality or was no longer engaged.

around the framework, and reflecting on its usability, credibility, and transparency (see supplementary file 1, section 1 for an explanation of the criteria). These criteria are key for developing a robust, systematic and practical evaluation (United Nations Evaluation Group, 2017).

Reflection by the evaluators was done continuously through bi-weekly meetings to discuss the progress and experiences. In a final evaluation meeting, the main experiences along the three evaluation criteria were discussed and summarized. Reflection by the stakeholders of the cases and the advisory group was done through two evaluation workshops. First, at a workshop of the project (June 2024, Gdynia), the framework in general and the case of Milan was presented, discussed, and evaluated. Because the city representatives of Lillestrøm could not join, their case was evaluated separately (September 2024, Lillestrøm). The evaluation workshops followed a similar structure, starting with a presentation on the framework and the outcomes of the cases. Thereafter, people were asked to reflect on it. Finally, all participants filled in a survey asking participants to reflect on the criteria of usability, credibility, and transparency. The outcomes of the discussions, surveys, and evaluation meetings were processed by summarizing the main outcomes and structuring them along the three evaluation criteria.

In preparing and carrying out the evaluation, the positionality of the evaluators was a key consideration, see also section 3.1.3 of supplementary file 2. Positionality refers to "*the recognition and declaration of one's own position in a piece of academic work*" (Rogers et al., 2013). Reflecting on the positionality of the evaluators is crucial for transparently communicating the potential influence on the evaluation process and outcomes. Here we briefly discuss the evaluators' position within the project and its implications for the research.

The evaluation was carried out by the first author (EB, PhD candidate and climate service advisor) and second author (NSB, junior researcher and consultant in the field of natural hazards). Both were also engaged in the project as climate story producers: EB for the city of Athens (not included in this paper) and NSB for Lillestrøm. Because of her role as producer of the Lillestrøm climate story, NSB had minimal engagement in its evaluation. The exception was the final evaluation workshop, which NSB facilitated to enable having it in Norwegian. The evaluation framework itself was developed by EB, with input from RB (third author), without engagement of any of the project partners.

As members of the project and producers of climate stories, both evaluators had a good understanding of the climate services, the cities, and the broader project. This provided them with easy access to interviews and relevant documents. Also, their expertise and experience in the field of adaptation and climate services allowed them to get a good understanding of the dynamics in the cases. Close engagement of evaluators can also threaten the impartiality and independence of the evaluation. Two main risks were identified at the start of the research. First, interviewees might feel pressured to evaluate the story positively, having to report to researchers that are part of the project themselves. To mitigate this, the learning goal of the evaluation was emphasized, and the stakeholders were encouraged to express their honest opinion. Second, as two of the authors of this paper contributed to the development of the climate story concept, they might themselves have bias towards positive outcomes. To mitigate this, preliminary ratings were discussed and checked with an external and independent researcher (third author, RB).

Evaluation framework for climate services

Here we present the design of the evaluation framework. Based on the literature review, 20 indicators were defined to measure the 12 success criteria, see Table 3. This is the final evaluation framework, including some minor changes based on testing and evaluating it (see sections 4 and 5). In this section we discuss the main considerations and decisions regarding the definition and selection of indicators.

Some criteria could clearly be measured well with a single indicator, either by investigating the success perception of stakeholders, or by making an informed judgement. The criteria *C1. Timely*, *C3. Credible*, *C4. Relevant*, and *R3. Benefits* could be directly measured through a single interview question. This was also the case for *R4. Relationship of Trust*, although here, trust as perceived by the users and by the producers are combined to evaluate the relationship of trust. An objective indicator could be defined to evaluate *C5. Uncertainty Communication*, by analysing the extent to which uncertainties were presented and discussed in the climate service.

For other criteria a more extensive approach was needed, combining two or more indicators, as well as using both success perception and objective evaluation. These indicators either need to *complement* each other or be *aligned*. For complementary indicators, the indicators measure different aspects that together provide a good understanding of the criterion. This is the case for *P1. Tailored Interaction* and *C6. Tailored Communication Format*. For both criteria there is rich literature on the need for tailoring, but there is no encompassing framework to guide the systematic evaluation of the best approach for different types of services, users, and contexts. Therefore, an indicator on the effort of tailoring is combined with an indicator on the outcome as perceived by the users. Another reason for using multiple complementary indicators was to make the measurement of a criterion manageable. The criterion *R5. Better Adaptation Decision-making* is difficult to measure due to problems with establishing causality and diverging perspectives on what is 'good' adaptation. To deal with this, the criterion was split up into different steps describing the potential to influence and improve adaptation decision-making.

In the second category, alignment, the indicators measure the criterion using different approaches that are partly overlapping. Each indicator has its own advantages and disadvantages, for example related to data collection and validity. The goal is therefore to achieve alignment between the indicators to be able to evaluate the criterion in a comprehensive and robust way. This is the case for criteria *C2. Accessible*, *R1. Increased Understanding,* and *R2. Increased Capacity*. Ideally, these criteria are measured objectively. This, however, requires extensive data which is often not available and may in many cases not be feasible to collect within a project context. Therefore, a strategy is suggested that allows the collection of evidence on the criteria (using 'light' methods) and combine this with perceived success of the targeted user group. For example, to measure the degree of accessibility (C2.1) a composite scale is provided combining 1) an estimation of the proportion of the target group that disposes of the resources and skills to access the climate service and 2) explore the extent to which (some) users from this group are able to achieve an intended goal with the service.

The selection of indicators was guided by the study's objective to develop a framework that allows *robust* and *systematic* evaluation, that is *practically* feasible in a project context (see section 1). This means that there were trade-offs in finding the best possible way to measure an indicator and dealing with time and budget limitations. We chose to limit data collection to (structured) interviews with users and producers, and desk study. Interviews are used to directly inquire on users' and producers' perceived success of a specific criterion, and to map and understand the development process, the climate service, and the results. If an indicator could be measured through a direct question in an interview, this was preferred over other more complex approaches where an indicator is reconstructed by the evaluator through analysing answers to multiple questions or by combining interviews with other methods (e.g. for criteria *C3. Credible* and *R4. Relationship of Trust*). Desk study may involve analysing the climate service itself, documents describing the service and/or its development process, correspondence between stakeholders, policy documents, and user statistics. The framework is flexible in how data can be collected, often proposing multiple ways that can be tailored to what is the most feasible and relevant approach for the case(s) under investigation.

To illustrate how the criteria are operationalized in the evaluation framework, we present the example *P1. Tailored Interaction*. This criterion assesses whether the interaction methods between the users and

Table 3
Indicators for measuring the success of climate services along 12 criteria. The success criteria were selected in a Delphi study where experts evaluated which elements are most relevant to define the success of climate services. Criteria from three conceptual categories were selected: 1) production process, 2) climate service, and 3) results of production and/or use. Additional criteria may be defined if relevant under the 'other' category. Supplementary file 2 presents further guidance on how the criteria can be measured (sections 2 and 3) as well as a justification for the selection of indicators (section 6.1).

Success criterion	Indicator
Production process	
P1. Interaction between users and producers is tailored to context	P1.1 Degree of tailoring efforts related to the interactions
The nature and frequency of interaction between producers and users – from highly collaborative to consultative – is tailored to the context (e.g. user and decision context).	P1.2 Degree of perceived suitability of the interactions
Climate service	
C1. Timely delivery	C1.1 Degree of perceived timeliness
The climate service is delivered in time to inform an intended decision or to satisfy a need in a specific timeline.	
C2. Accessible climate service	C2.1 Degree of accessibility
Users can access, interact with, and understand the climate service.	C2.2 Degree of perceived accessibility
C3. Credible information	C3.1 Degree of perceived credibility
Users perceive the information in the climate service as reliable and trustworthy.	
C4. Relevant information	C4.1 Degree of perceived relevance
Users perceive the information in the climate service as relevant to their needs, problems and/or decision-making.	
C5. Acknowledgement and communication of uncertainty	C5.1 Degree of uncertainty communication
The climate service acknowledges and communicates the uncertainty associated with climate change information.	
C6. Communication format is tailored to users	C6.1 Degree of tailoring efforts related to the communication format
The climate service communication format and messaging strategies are tailored to the users and their needs, think of using appropriate language and suitable media.	C6.2 Degree of perceived suitability of the communication format
Results of production and/or use	
R1. Climate service increases users' understanding of an issue	R1.1 Increase of understanding of an issue
The climate service increases the users' understanding of an issue. For example: users may feel better informed on future impacts or are capable to reframe the problem and identify possible solutions.	R1.2 Degree of perceived increase of understanding of an issue
R2. Users build the capacity for using services	R2.1 Increase of user capacities
Users learn how they can use the climate service and how it may benefit their decision-making, which in turn may drive future demand for services.	R2.2 Degree of perceived learning by users
R3. The climate service has tangible or intangible benefits for the user	R3.1 Degree of perceived degree of benefits by users
Think of feeling more safe, being better prepared or an increase of income, employment, or literacy.	
R4. Establishment of trust between users and producers	R4.1 Degree of trusted relationship – perceived by users
The users and producers of a climate service establish a relationship of trust	R4.2 Degree of trusted relationship – perceived by producers
R5. Better decision-making for adaptation	R5.1 Potential of the climate service influencing adaptation decision-making
The climate service contributes to better decision-making for adaptation, e.g. through informing policies or actions that decrease climate vulnerability or improve adaptive capacity.	R5.2 Degree of perceived increase of decision-making capacity
	R5.3 Potential of climate service for supporting adaptation/maladaptation
Other (optional; here an example from the testcases is presented)	
O1. Fair process	O1.1 Degree of efforts for engaging relevant stakeholders
All relevant stakeholders were engaged in the development process of the climate story and their knowledge was integrated in an appropriate way.	O1.2 Degree of perceived legitimacy

producers are tailored to the specific context in which the climate service is produced. Here, 'interaction method' refers to the nature and frequency of interactions, which may range from highly collaborative to more consultative, occur regularly or occasionally, and take place face-to-face, online and/or through written correspondence. The most suitable method may depend on the users' needs, capacities, and decision-context (e.g. see Lemos et al., 2019; Meadow et al., 2015). For example, in-person interaction may be more suitable when the climate service involves more complex climate change information or when there are trust issues regarding the use of science (Lemos et al., 2019). To clarify, this criterion does not favor close and frequent interactions as is advocated in many climate service publications and guidelines. Instead, it evaluates whether the chosen interaction method is appropriate for the given context. Currently, there is no encompassing framework or evidence base outlining which types of interactions are most effective, under what circumstances, and for which users, to guide the assessment of this criterion. As an alternative, we assess this criterion using two complementary indicators:

1. Degree of tailoring efforts related to the interactions (P1.1): This indicator reviews the extent to which alternative interaction methods were considered and whether the selected method was validated to meet the users' needs and their decision-context.
2. Degree of perceived suitability of the interaction (P1.2): This indicator assesses the extent to which the 'tailoring' of interactions succeeded, as perceived by the targeted users.

A detailed explanation and justification for the operationalization of all 20 criteria can be found in Tables 1 and 2 of supplementary file 2. This includes detailed metrics describing how to evaluate an indicator on a rubric scale from 1 to 5 as well as directions for data collection and analysis. The same document includes the protocol for evaluation (section 3). This protocol is an important result as it describes key process steps for applying the evaluation framework, such as setting the boundaries of the evaluation, preparing the evaluation with a data collection plan and evaluation materials, and directions for processing, structuring and reporting on the collected data. This includes, for example, how indicators can be combined into a final rating for the criterion and what to do if no data is available. It also describes how a robustness score for each criterion can be established, by reflecting on the representativeness of the sample and the extent to which different sources could be used to validate information. A qualitative approach is suggested, providing either a low, medium, or high robustness score. Finally, in addition to the protocol, a checklist and set of guiding questions were formulated to support integrating evaluation in the development stages of a climate service (see section 4 in supplementary

file 2). This guidance was developed in response to the experience from the evaluators and case stakeholders that opportunities were missed for developing successful climate stories and collecting data for evaluation (see also section 5 of this paper), as well as to provide an easy-to-understand overview of the framework.

Evaluation of the climate stories of Milan and Lillestrøm

The evaluation framework was tested in two case studies. As explained in section 2.3, '*O1. Fair Process*' was included as an extra evaluation criterion. In this section, we provide a brief overview of the cases and the outcomes of testing the framework. Additionally, we highlight key strengths and weaknesses identified by the framework to demonstrate the type of insights it can offer. The detailed evaluation of the cases, along with its justification, is presented in supplementary file 1, sections 4 and 5.

Case 1: Milan climate story

Ambrogio and Gaia – A climate story about Milan's heatwaves

About 1.3 million people live in the municipality of Milan (Italy), and an estimated 3 million when including the wider metropolitan area. The city considers heatwaves as a big threat, especially to citizens in the most built-up areas. Milan therefore aims to increase green space in the city and develop a heat strategy as part of their Air and Climate Plan (Direzione Transizione Ambientale, 2022). The city participates in various European projects to support their efforts on adaptation, mitigation, and resilience, and has previous experience with developing and using climate services. When storytelling was offered as a climate service in the project, the city representatives recognized it as a promising instrument to reach and engage the general public. They aimed to use it to inform citizens about future heat impacts, the importance of green solutions, ongoing municipal plans and actions, and opportunities for community involvement.

In the summer of 2022, the first ideas for the content of the story were developed by the producers in a workshop. Next, at a workshop (Milan, October 2022), the building blocks of the story were further co-produced with various city representatives of Milan, as well as city representatives from the cities of Logroño and Athens, who also face challenges regarding heat. After that, first drafts of the story were developed and shared with the city representatives of Milan. The main city representative followed up internally in the city administration to collect input and get approval for the climate story. The story for Milan was developed in a relatively direct process, involving some iterations with the producers to make changes to the content. Citizens were not involved in the development process.

The story was finished early 2023 and is available in English and Italian.[1] The story follows Ambrogio and his granddaughter Gaia on their way to the park during a summer heatwave. The story presents data on historic and future heatwaves through graphs and infographics. The concept of 'generations' is used to link heat wave frequency and intensity over time, to the lives of Ambrogio and Gaia. The goal of this was to make the citizens connect these climate statistics to their own lives.

Various city stakeholders presented the story at events and meetings throughout the year 2023 (e.g. Green Week, at universities). It was not publicly launched. While using the story at events, city stakeholders gained new insights into how they could utilize and expand the story's impact. These ideas emerged from their own experiences and from feedback they received. This led to the city representative deciding to engage a climate illustrator to improve the pictures in the story and further contextualize the narrative. The idea is to develop a physical booklet of the story that you can fold out into a poster ('fanzine') with QR codes to view and inspect maps in the 'online' version of the story. This happened mid 2024 – just before the end point of the evaluation. In addition, the city representative started a process to gather input on the content of the story from the general public through the climate citizen assembly – a heterogenous group of about 100 citizens.

Evaluation of the Milan climate story

Fig. 2 summarizes the evaluation of the Milan climate story. The process-related criteria fluctuate around 3 (on a scale of 1–5), being evaluated as neither successful nor unsuccessful. Most of the criteria related to the product and the results fluctuate around the rating 4 (towards more successful). Criterion *C5. Uncertainty Communication* is recognized as an outlier, as it was evaluated with a 1: unsuccessful.

Most of the evaluated success criteria score a low on their robustness. This is because many criteria involve indicators that measure the success perception of the targeted users (citizens in this case), but they were not engaged in developing the story. The lack of interaction channels hampered engaging them in evaluation. As an alternative strategy, data for these indicators was collected by interviewing the various city representatives engaged in the story that are also inhabitants of the city. They, however, cannot be considered representative for all citizens of Milan. The scores of the process criteria were also directly influenced by the lack of citizen engagement as they involve indicators that measure efforts to engage relevant stakeholders (O1.1) and apply suitable interaction methods (P1.1).

For four out of 20 indicators (C2.1, R1.1, R2.1 and R5.1) there was insufficient data to provide a rating. This was mainly because there were no clearly defined goals and baselines, and there was no data or possibilities to collect data about potential impacts of the story on citizens. These indicators were excluded from the rating of the criterion, further compromising the robustness scores of the success criterion.

The evaluation revealed that not engaging citizens in the development and evaluation of the story led to missed opportunities to discuss, validate, and evaluate the development process, the climate story itself, and potential impacts. The city representative reflected that more people, including citizens, could have been engaged in the process. At the start of the process, the idea for the story was still very unclear and it wasn't evident who to engage. The city representative therefore preferred to keep the process simple. Despite the lack of citizen engagement and measurement of citizen related impacts, various positive outcomes were reported by and observed for the city stakeholders themselves, including developing a better capacity for developing stories and knowing when and how they could be used. This learning informed the decision to organize citizen consultations to further develop the story, representing a promising opportunity to improve the climate story and support future evaluation efforts. The framework also highlighted the story's potential for informing adaptation decision-making. The city representative highlighted that the story was helpful in achieving goals of the Air and Climate Plan, aligning objectives across various departments (such as urban resilience, welfare, and health), and demonstrating the benefits of participating in European projects to colleagues. An improvement highlighted by the evaluation is to include information in the story about the selection and meaning of RCP 8.5 scenario in presenting heatwave data, as this would increase the score of the "Uncertainty communication" indicator.

Case 2: Lillestrøm climate story

Life by and with water: Many small streams – A story about climate adaptation in Lillestrøm

Lillestrøm municipality (Norway) has around 95.000 inhabitants. The city is located close to the capital city, Oslo, and has an important economic and social function in the region, providing housing, employment, and an extensive infrastructure network. The urban area is expanding and will be further developed in the coming decades. Being part of an inland delta, the city has a history shaped by flood events,

[1] https://reachout-cities.eu/climate-stories/.

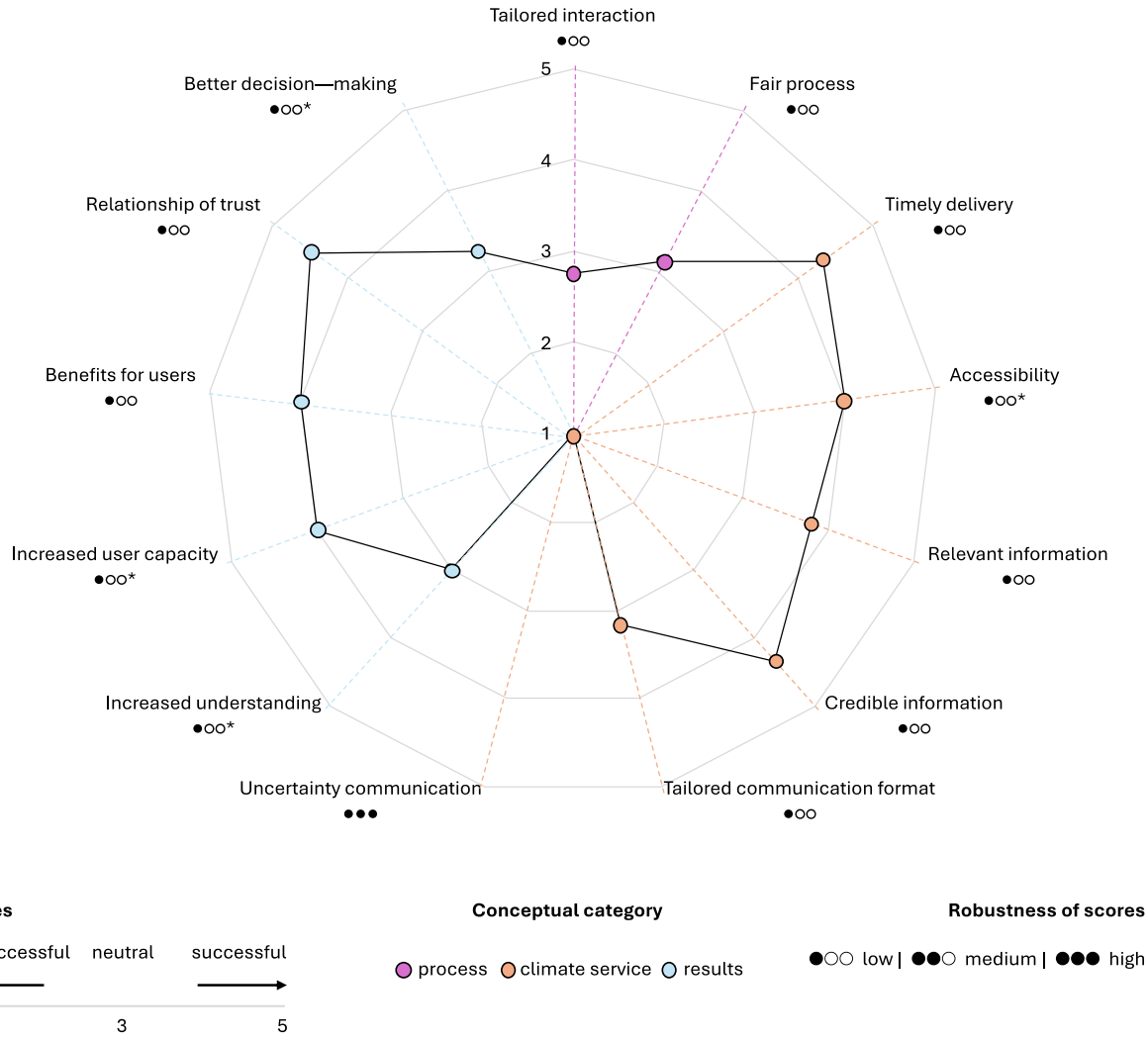


Fig. 2. Milan climate story evaluation along 13 success criteria. '*' indicates that no data was available for one of the indicators of the criterion.

both caused by the river and extreme rainfall. Levees and pumps now protect the city from most river floods, however the city recognizes that changing precipitation patterns combined with increasing urbanization increases flood risks that cannot be mitigated by grey measures alone. They are therefore seeking ways to implement nature-based solutions and want to engage citizens and stakeholders working construction and development ('builders') to achieve this. The city representatives recognized that a climate story could help them create awareness about climate change impacts due to pluvial flooding, possible solutions, as well as encourage citizens to take action. The targeted user groups are citizens, especially homeowners, people working for the municipality, and builders.

The climate story was developed by the city representatives and the producers in an iterative process, starting with the identification of 'building blocks' for the story ('intake meeting', Summer 2022). This was followed by an internal brainstorm between the producers of all climate story teams to think of possible storylines. Thereafter there were online meetings and emails between the producers and the city representatives to discuss draft story versions, as well as two workshops in the municipality to gather feedback and input from a larger group of city stakeholders. Citizens and builders were not engaged in this process.

The story was finished early 2024 and is available in Norwegian and English.[2] In the story, you follow Sofie and her grandmother Kari on their weekly walk through the city. They notice something different from usual: the water in the river is higher and flowing faster than usual. Kari discusses the city's history of flooding as they walk past various solutions the municipality have implemented throughout Lillestrøm's city center. Additionally, the story introduces various adaptation measures that citizens themselves can implement to mitigate flood impacts, and explain how they work.

The story was published online and publicized through news items on the municipal website,[3] LinkedIn, and Facebook, and through digital posters with QR codes throughout the city. It was also promoted in a local newspaper interview with one of the city representatives. In the period between the launch and the end point of the evaluation there was a reorganization in the municipality, redistributing adaptation responsibilities. As a result, the climate story was transferred to another city representative and a new department.

Evaluation of the Lillestrøm climate story

Fig. 3 summarizes the evaluation of the Lillestrøm climate story. It shows a similar pattern as the evaluation of the Milan story: the process

[2] https://reachout-cities.eu/climate-stories/.
[3] https://www.lillestrom.kommune.no/aktuelt/nyhetsarkiv/2024/klimafortellingen-om-lillestrom/.

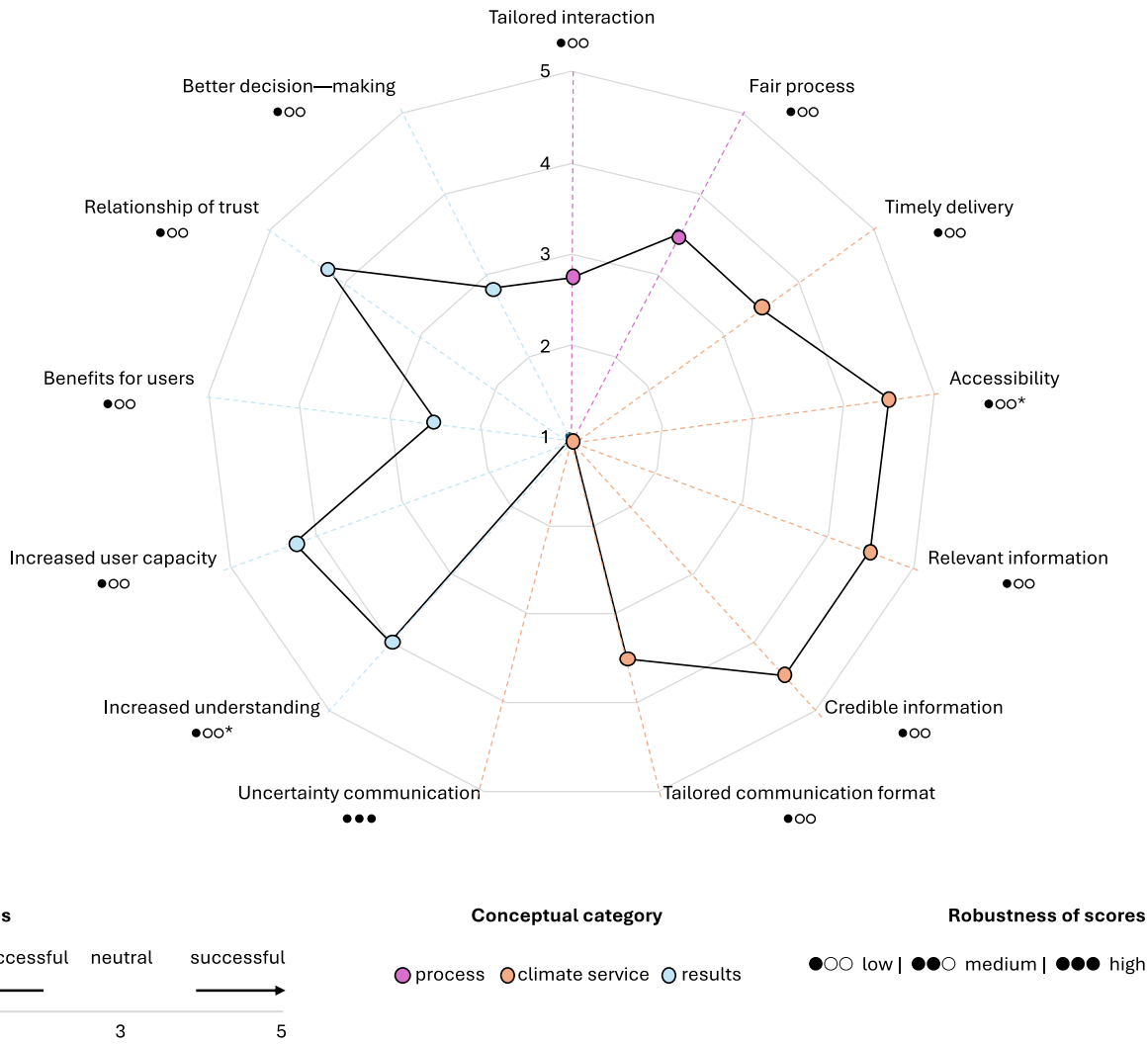


Fig. 3. Lillestrøm climate story evaluation along 13 success criteria. '*' indicates that no data was available for one of the indicators of the criterion.

related criteria fluctuate around 3, and the product and result related criteria fluctuate around 4. Also here, *C5. Uncertainty Communication* scores 1: unsuccessful. A difference compared to Milan's story is that *R3. Benefits for the users* is evaluated with 2.5.

Similarly to Milan, the robustness score of most success criteria is low. This is because two of the three targeted user groups were not involved in developing and evaluating the story, while many criteria involve indicators measured through user perception. As explained with the Milan story evaluation, the lack of user engagement in this process directly influences the process related criteria.

For two out of 20 indicators (indicators C2.1 and R1.1) there was insufficient data to provide a rating. This was mainly because there were no clearly defined goals and baselines, and there was too little data about the impacts of the story on the user groups.

Like with the Milan case, including all targeted user groups in the development process would have provided better opportunities to discuss, validate, and evaluate the development process, the climate story itself, and potential impacts. The city stakeholder wasn't sure yet how and when they could use and exploit the story, apart from launching it on their website. However, the evaluation revealed that the city stakeholders appreciated the process and output, as they developed a better understanding of how to communicate to citizens and builders in an easy-to-understand way. The city representatives reflected that it would be good to engage citizens and builders in future work on the story. This learning motivates the city stakeholders to continue their work on stories, including developing stories for other topics (e.g. biodiversity). Another benefit identified by the framework, was that developing the story with stakeholders from different departments strengthened their collaboration on adaptation topics. Finally, the evaluation revealed potential for improvement by discussing the uncertain nature of precipitation patterns and future flood risk.

Evaluation of the evaluation framework

Table 4 summarizes how the usability, credibility, and transparency of the framework was evaluated by the city representatives of the cases, the advisory group, and by the evaluators themselves. This is further detailed in the next sections as well as revisions to the framework based on the evaluations.

Usability

In general, the evaluation framework was considered usable for a variety of applications in research and practice, such as including it in the terms of reference of calls for tenders, developing business models, use it as design criteria, guide development processes, support monitoring and evaluation, facilitate learning about what works and what doesn't work, and provide a framework for reporting. Especially the stakeholders from the 'producing side' of the advisory group valued the framework. For the city representatives of Milan, the evaluation of their

Table 4
Summary of the evaluation of the evaluation framework, indicating whether the criteria were evaluated as high (+), medium (+/-), or low (−), and if recommendations were provided (>).

	Usability	Credibility	Transparency
City representatives of Milan	+ Use as design criteria for next story. − No new learnings, improvement plan was already established.	+ Evaluation outcomes were recognized.	/
City representatives of Lillestrøm	+ Useful to reflect on your investments. + Guide with success criteria and might be useful for developing future stories. + Learning: better and wider involvement of users and setting measurable goals. > In the communication of the results (i.e. spider diagram), specify the goals of the climate story.	+ Evaluation outcomes were recognized. − Limited validity: users engaged in evaluation not representative for target audience.	+/- Unsure who exactly were involved in the evaluation.
Advisory group	+ Use in/as calls for tenders, design criteria, business models, development, monitoring, evaluation, and learning. Use by tool and service developers, and knowledge brokers. + Relevant criteria, including those that otherwise might be overlooked (e.g. tailored interaction and intangible benefits). − Using all criteria for all types of services. > Select criteria upfront. Test guidance. Promote uptake of the framework in the climate service community.	+ Clear indicators, based on credible sources. Combining subjective and objective indicators. Option to add criteria. + Using mixed methods. + Selecting suitable evaluator. +/- Framework needs further testing. − Missing criteria; sustainability, marketability, ownership, and data quality.	+ Clearly structured framework. +/- Comprehensive, difficult to understand quickly. > Develop one-pager for quick overview.
Evaluators	+ Clear indicators, metrics and methods are usable to rate the criteria consistently. +/- Time consuming. Data collection challenges. Indicator focused. > Integrate evaluation in development.	+ A clear framework with detailed instructions and protocols, along with checks by an external researcher, ensured an objective evaluation. Indicating robustness was highly important. − Data collection challenges.	+ Transparent communication on data robustness. + Providing justification for each of the ratings.

case did not reveal much practical relevance, as they had already identified learning points themselves. Using design criteria as guidance to develop future stories sparked more interest. For the city representatives of Lillestrøm, the evaluation of their case had more practical value. They identified two main lessons: better engaging users during the development and evaluation of the story and setting clear and measurable goals.

In the advisory group a discussion emerged on whether all criteria should always be evaluated. Some people thought that the low rating for uncertainty communication in the Milan case was wrong, reasoning that this criterion is not relevant when the main goal of the service is to increase awareness of citizens. Instead, they suggested to select criteria relevant to a case upfront of the evaluation, with a proper justification. City representatives of the Lillestrøm were unsure what would be a good way to communicate uncertainty.

In the evaluators' experience, the framework was usable to rate each indicator consistently. Retrieving success perceptions through statements in interviews worked well. People were confident in providing ratings as well as explanatory comments and context. Based on the experiences of the evaluators, minor changes were made to the initial evaluation framework. For example, the evaluators originally tried measuring *C5. Uncertainty Communication* by considering users' perception of this criterion, alongside an objective evaluation. However, this approach didn't provide any useful insights. The question (tested in several forms) often caused confusion, and after clarification, users usually didn't have a clear opinion on the matter. The clearly defined metrics were helpful for the indicators that require an informed judgement. There were also some challenges with applying the framework. Some data was not available and difficult to reconstruct or time consuming. For example, because the user groups and the goal(s) of the climate stories were defined in very general terms, it was difficult to get a grip on criteria from the 'results' category, and design appropriate strategies to collect evidence. In addition, as evaluator, you need time to familiarize yourself with the framework to be able to collect data in an efficient and integrated way for 20 indicators. Finally, to keep the evaluation manageable, there is a need to keep a certain focus on measuring the indicators, risking that other experiences and developments that are relevant for the goal of the evaluation are missed out on.

Credibility

Overall, the evaluation framework was considered credible because it is based on credible sources and uses a rigorous methodology, including providing a robustness score, using mixed methods, combining objective with subjective indicators, and deliberately selecting a suitable evaluator. This was further supported by the fact that the stakeholders of the evaluated cases acknowledged and agreed with the evaluation outcomes. City representatives of the Lillestrøm case however, recognized and indicated that the outcomes have limited validity since it didn't include a representative group of users.

While the criteria used in the framework were generally considered comprehensive, some people from the advisory group thought one or two criteria were missing. For example, some climate service tool developers, emphasized that data quality should be included. They recognized that it may be included indirectly in the criteria about trust and credibility, but in their perspective the scientific perception of credibility was missing. In this context it was appreciated that an additional criterion could be added to the evaluation. Furthermore, people from the advisory group agreed that the framework needs to be tested in different types of services to further enhance credibility.

From the evaluators' perspective, the main challenge was to develop reliable ratings for the cases. In both cities it was very difficult to reach a representative group of users to inquire on their success perceptions. Also, the lack of clearly defined goals discussed above made it difficult to collect evidence for some indicators. The predefined criteria and metrics were very important to develop objective ratings. For some criteria, the ratings based on the framework didn't immediately align well with the evaluator's 'intuition', and it was helpful to discuss it with an independent researcher. For example, from the interviews with various stakeholders, it emerged that several valuable capacities were developed by the city representatives, such as being able to develop stories themselves and having a better understanding of what climate services they need to support their adaptation efforts. However, there was often little data (e.g. lack of baseline, based on only one source, and not representative for the user group) to develop a robust rating for this criterion. The robustness score appeared to be a crucial aspect to be able to communicate about the trustworthiness of the rating.

Transparency

The transparency of the evaluation framework was considered adequate as all evaluation materials were accessible and well-structured. However, because the framework and protocol are rather comprehensive, it was recommended by the advisory group to develop a 1-pager in which some concepts and the main approach are described concisely, to further increase the transparency and accessibility. This should include, among other things, that the framework addresses the service, its development process, as well as impacts, and clearly explain the criteria with an indication of rating 1–5, the term user, and the term climate service. The stakeholders of the cases showed limited interested in how the ratings were established or how the framework was applied. The spider diagram and a summary of the evaluation was sufficient for them to discuss the outcomes, but it could have been clearer for the city representatives who had been engaged in the evaluation.

Discussion

In this discussion, we reflect on the development (Section 3), testing (Section 4), and evaluation (Section 5) of the evaluation framework, sharing lessons learned, and suggesting areas for future research.

An operational evaluation framework to progress research and practice

This paper reports on our study that designed, tested, and evaluated an evaluation framework for climate services for adaptation. The framework contributes to closing the gap in climate services literature about how their success can be evaluated (Boon et al., 2022; Englund et al., 2022). In doing so, the framework offers a standardized approach to evaluation, providing clear indicators and metrics for agreed upon success criteria. By using mixed methods, combining objective and subjective indicators, applying standardized interviews questions, and critically assessing the robustness of the evaluation, the framework offers a strong assessment of climate service success, enabling their systematic evaluation and comparison. The framework can be applied to the broad range of climate services and contexts, by tailoring metrics and choosing a feasible approach for collecting data. All together, the framework contributes to the standardization of climate services, which is called for by various authors and climate service projects to mitigate inequalities in the quality of climate services and prevent misguided adaptation decisions and decreased trust in science and derived services (Baldissera Pacchetti and St.Clair, 2023; Guentchev et al., 2023).

The framework is usable for both researchers and practitioners, for a variety of applications. While generally perceived as credible and transparent, the framework would benefit from further testing and elaboration. A first step is to test the framework with other types of users and climate services, such as data-intensive (e.g. data tools) or process-intensive services (e.g. workshop methods) and to explore the value of various potential applications. This also involves the exploration of climate services that engage users throughout their development process. Second, two out of twenty indicators (C2.1 and R1.1) couldn't be tested in the cases and require further exploration and validation. Additionally, the indicators designed to objectively measure intangible benefits, such as increased capacity, are challenging to assess robustly and would benefit from additional testing. Third, the guidance developed to consider success criteria from the start of the development process of a climate service, needs to be applied in practice to find out if it indeed is as helpful as expected. Fourth, as we gain more experience with applying the framework and scientific knowledge advances, it may be necessary to update the indicators with more appropriate alternatives. For example, a deeper understanding of the types of interactions (criterion P1) or communication formats (criterion C6) suitable for different services and contexts could allow current indirect indicators to be replaced with more direct measures. Additionally, it is important to reflect on the appropriateness of using subjective and/or objective indicators for each criterion.

Good evaluation starts before development

The results of this study demonstrate once again that good evaluation is done best when it is integrated early in the development process of a climate service. First, it allows more efficient data collection, by being able to link data collection to development milestones and setting up suitable interaction channels with users. In the evaluated cases, these interaction channels were poorly established or absent, hindering the evaluation of some indicators. In addition, early integration of evaluation provides opportunities to minimize the risk of stakeholder fatigue, by eliminating 'extra' evaluation moments. Second, early consideration of the criteria allows collecting data that is difficult or impossible to reconstruct at a later stage. This can, for example, apply to defining and delineating the targeted users and the desired goals of the service in a suitable way, including establishing baselines and defining what counts as evidence to be able to find any 'change'. This is in line with various studies that suggest developing theories of change beforehand, to improve the development and evaluation of services by specifying how the service is intended to cause change and lead to outputs and impacts (Englund et al., 2022; Kalsnes et al., 2023; Tall et al., 2018). The evaluation framework effectively highlights the poor definition of the targeted user groups and goals in the studied cases. Users and goals were described in general terms, making it challenging to reconstruct a baseline and measure any impact. Moreover, we recognized that the cases have a chain of users, where those engaged in the development (city representatives) are not the targeted 'end-users' (mainly citizens). This means that benefits of engagement (e.g. increased trust, understanding about what is relevant, credible, accessible for the users, increased capacities) may end up with the 'wrong' people. Hence, in such models, it is necessary that relevant goals and baselines are specified for the respective groups.

An additional and important benefit of integrating evaluation early in development, is that it enables learning and adjustment (Englund et al., 2022), potentially leading to the development of more successful services. By considering evaluation from the start, producers (and possibly users) are aware of success criteria that otherwise may be overlooked, and they can change the direction of development accordingly. For example, in the cases, the communication of uncertainty was scarcely considered or discussed during the development of the stories. This might have been different if it had been emphasized as a success criterion. There are still few examples of evaluation being integrated in and informing development, but in cases where it has been integrated, researchers found that it led to more iterative processes and better engagement of stakeholders (Kalsnes et al., 2023). This in turn may lead to more successful services, as it could improve tailoring and inclusion of user perceptions, which are central factors for the success of climate services.

Beyond a snapshot of success

The results of applying the evaluation framework should be seen as a snapshot of success, and has most value if it is understood in its wider context and updated over time. Low success ratings and robustness scores do not mean that a service cannot develop into a successful service, or that the evaluation is useless. Contrastingly, cases where learning is lacking are cases where services are more likely to fail. The climate stories evaluated in this paper provide a good example of this. Due to the limited understanding and intent regarding the specific goals and targeted'end-users' of the story, some indicators couldn't be measured and the robustness scores of many evaluated criteria remained low. The cases are not unique in this. For years, literature has reported on services that are overly science-driven and lack user engagement and customization (Findlater et al., 2021; Lemos et al., 2012; Weichselgartner and Arheimer, 2019). To achieve the opposite, demand-driven

development is advocated (ibid). Such processes, however, may indeed require that users and goals remain open early in the process, and only become better defined after several iterations of development and application, through which learning occurs. This was clearly visible in the Milan case where the city representatives only developed a good understanding of how they could and wanted to use the story, after a year of experimentation. This led to the development of an improvement plan, including engaging citizens to further tailor and contextualize the story. This implies that the comprehensiveness and robustness of the evaluation, as well as the success of the service itself, is likely to improve over time if, through learning and user consultation, the users and goals are further specified.

Balancing flexible with robust and systematic evaluation

Whether all success criteria should be evaluated for all types of climate services was a central point of discussion in the evaluation of the framework. Arguing that some criteria are more relevant for certain types of services than others (see also Bremer et al., 2021), it was suggested to be flexible in the use of the criteria or to weigh the criteria differently. However, we advise against discarding criteria that evaluators, producers, or users consider irrelevant on a case-by-case basis, prompted by feelings, experience, or preference. The added value of the standardization of an evaluation framework is precisely that it builds on a set of agreed success criteria which are applicable to the broad range of services (Boon et al., 2024). Adding one or two criteria, as we did for the two cases here, rather than removing existing ones, could reconcile this. Of course, evaluators should follow the framework and guidelines as closely as possible but should also remain vigilant about whether the indicators and metrics continue to effectively measure the criteria.

Fostering the uptake and application of the evaluation framework

The study identified a diversity of framework applications which were considered usable and valuable for advancing climate service research and practice. Given the lack of an evaluation culture, however, we see a challenge for the uptake of the framework in the climate services community. Evaluation may be constrained by the direct costs and lack of perceived benefits. Typically, it is seen as technical requirement, rather than a process that is inspiring, fun, and helpful. We therefore think that the framing and presentation of the framework matter. To stimulate uptake and application, we see two promising pathways. First, there is an opportunity to connect evaluation to the practice where climate services are developed in multiple development cycles while engaging users. The criteria identified in this framework could structure and guide this process. By promoting the direct added value of the framework in development processes, awareness on the criteria could be increased, paving the way for more systematic efforts to evaluate them. A second pathway is to focus on the necessity of evaluation, for example to be able to mitigate misguiding or ineffective services. Commissioning parties could demand evaluation and reporting in their calls for tenders or include it as design criteria in the terms of reference. In either case, the framework, together with the protocol and guidance, contributes to the much-needed toolbox for enabling robust and systematic evaluation of climate services.

CRediT authorship contribution statement

Eva Boon: Writing – original draft, Methodology, Investigation, Formal analysis, Conceptualization. **Nellie Sofie Body:** Writing – review & editing, Validation, Investigation, Formal analysis. **Robbert Biesbroek:** Writing – review & editing, Supervision, Methodology, Conceptualization.

Funding

This research took place in the context of, and was funded by, the REACHOUT project. The REACHOUT project has received funding from the European Union's Horizon 2020 Research and Innovation Programme under Grant agreement no. 101036599. The authors had cart blanche to perform this research.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgement

We would like to thank all partners of the REACHOUT project for contributing with their views and experiences in the various interviews and workshops related to the climate stories and the evaluation framework developed and tested in this study.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.cliser.2025.100549.

Data availability

The data that has been used is confidential.

References

André, K., Järnberg, L., Gerger Swartling, Å., Berg, P., Segersson, D., Amorim, J.H., Strömbäck, L., 2021. Assessing the Quality of Knowledge for Adaptation–Experiences From Co-designing Climate Services in Sweden. Front. Clim. 3, 1–12. https://doi.org/10.3389/fclim.2021.636069.

Baldissera Pacchetti, M., St.Clair, A.L., 2023. Framework to support the equitable standardisation of climate services, D1.2 of the Climateurope2 project.

Barel, A., 2020. Storytelling en de wereld. International Theatre & Film Books.

Belcher, B.M., Rasmussen, K.E., Kemshaw, M.R., Zornes, D.A., 2016. Defining and assessing research quality in a transdisciplinary context. Res. Eval. 25, 1–17. https://doi.org/10.1093/reseval/rvv025.

Boon, E., Meijering, J.V., Biesbroek, R., Ludwig, F., 2024. Defining successful climate services for adaptation with experts. Environ. Sci. Policy 152, 103641. https://doi.org/10.1016/j.envsci.2023.103641.

Boon, E., Wright, S.J., Biesbroek, R., Goosen, H., Ludwig, F., 2022. Successful climate services for adaptation: What we know, don't know and need to know. Clim. Serv. 27, 100314. https://doi.org/10.1016/j.cliser.2022.100314.

Bremer, S., Wardekker, A., Jensen, E.S., van der Sluijs, J.P., 2021. Quality Assessment in Co-developing Climate Services in Norway and the Netherlands. Front. Clim. 3, 1–15. https://doi.org/10.3389/fclim.2021.627665.

Direzione Transizione Ambientale, 2022. Piano AriaClima. Milano.

Englund, M., André, K., Gerger Swartling, Å., Iao-Jörgensen, J., 2022. Four Methodological Guidelines to Evaluate the Research Impact of Co-produced Climate Services. Front. Clim. 4. https://doi.org/10.3389/fclim.2022.909422.

Findlater, K., Webber, S., Kandlikar, M., Donner, S., 2021. Climate services promise better decisions but mainly focus on better data. Nat. Clim. Chang. 11, 731–737. https://doi.org/10.1038/s41558-021-01125-3.

Guentchev, G., Palin, E.J., Lowe, J.A., Harrison, M., 2023. Upscaling of climate services – What is it? A Literature Review. Clim. Serv. 30, 100352. https://doi.org/10.1016/j.cliser.2023.100352.

IPCC, 2022. IPCC AR6 Working Group II: Summary for policymakers: Climate Change 2022, Impacts, Adaptation anf Vulnerability, in: Contribution of Working Group II to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change. pp. xxiii–xxxiii.

Jahan, M., Reis, J., Shortridge, J., 2023. Assessing climate service products with evaluation metrics: an application to decision support tools for climate change adaptation in the USA. Clim. Change 176, 113. https://doi.org/10.1007/s10584-023-03595-0.

Kalsnes, B., Oen, A., Frauenfelder, R., Heggelund, I., Vasbotten, M., Vollstedt, B., Koerth, J., Vafeidis, N., van Well, L., Jan Ellen, G., Koers, G., Raaphorst, K., 2023. Stakeholder evaluation of the co-production process of climate services. Experiences from two case studies in Larvik (Norway) and Flensburg (Germany). Clim. Serv. 32, 100409. https://doi.org/10.1016/j.cliser.2023.100409.

Larosa, F., Mysiak, J., 2019. Mapping the landscape of climate services. Environ. Res. Lett. 14, 093006. https://doi.org/10.1088/1748-9326/ab304d.

Lemos, C.M., Kirchhoff, J.C., Ramprasad, V., 2012. Narrowing the climate information usability gap. Nat. Clim. Chang. 2, 789.

Lemos, M.C., Wolske, K.S., Rasmussen, L.V., Arnott, J.C., Kalcic, M., Kirchhoff, C.J., 2019. The Closer, the Better? Untangling Scientist-Practitioner Engagement, Interaction, and Knowledge Use. Weather Clim. Soc. 11, 535–548. https://doi.org/10.1175/WCAS-D-18-0075.1.

Meadow, A.M., Ferguson, D.B., Guido, Z., Horangic, A., Owen, G., Wall, T., 2015. Moving toward the Deliberate Coproduction of Climate Science Knowledge. Weather. Clim. Soc. 7, 179–191. https://doi.org/10.1175/WCAS-D-14-00050.1.

Rogers, A., Noel, C., Rob, K., 2013. Oxford Reference: Positionality [WWW Document]. A Dict. Hum, Geogr https://www-oxfordreference-com.ezproxy.library.wur.nl/display/10.1093/acref/9780199599868.001.0001/acref-9780199599868-e-1432 (accessed 8.20.24.

Schuck-Zöller, S., Bathiany, S., Dressel, M., El Zohbi, J., Keup-Thiel, E., Rechid, D., Mirko, S., 2022. Developing criteria of successful processes in co-creative research. A Formative Evaluation Scheme for Climate Services. https://doi.org/10.22163/fteval.2022.541.

Street, R.B., 2016. Towards a leading role on climate services in Europe: A research and innovation roadmap. Clim. Serv. 1, 2–5. https://doi.org/10.1016/j.cliser.2015.12.001.

Tall, A., Coulibaly, J.Y., Diop, M., 2018. Do climate services make a difference? A review of evaluation methodologies and practices to assess the value of climate information services for farmers: Implications for Africa. Clim. Serv. 11, 1–12. https://doi.org/10.1016/j.cliser.2018.06.001.

United Nations Evaluation Group Norms and Standards for Evaluation 2017 New York.

Vaughan, C., Dessai, S., 2014. Climate services for society: Origins, institutional arrangements, and design elements for an evaluation framework. Wiley Interdiscip. Rev. Clim. Chang. 5, 587–603. https://doi.org/10.1002/wcc.290.

Visman, E., Vincent, K., Steynor, A., Karani, I., Mwangi, E., 2022. Defining metrics for monitoring and evaluating the impact of co-production in climate services. Clim. Serv. 26, 100297. https://doi.org/10.1016/j.cliser.2022.100297.

Wall, T.U., Meadow, A.M., Horganic, A., 2017. Developing Evaluation Indicators to Improve the Process of Coproducing Usable Climate Science. Weather Clim. Soc. 9, 95–107. https://doi.org/10.1175/WCAS-D-16-0008.1.

Walter, A.I., Helgenberger, S., Wiek, A., Scholz, R.W., 2007. Measuring societal effects of transdisciplinary research projects: Design and application of an evaluation method. Eval. Program Plann. 30, 325–338. https://doi.org/10.1016/j.evalprogplan.2007.08.002.

Weichselgartner, J., Arheimer, B., 2019. Evolving Climate Services into Knowledge-Action Systems. Weather. Clim. Soc. 11, 385–399. https://doi.org/10.1175/WCAS-D-18-0087.1.