Project Number 700099
Call: H2020-DRS-01-2015

Project Title:

ANYWHERE

EnhANcing emergency management and response to extreme WeatHER and climate Events

Subject:

Deliverable 1.3:
Report describing the methodology for the co-evaluation of the ANYWHERE distributed platform For Decision support in Emergency Management Operation Services (A4DEMOS)

Dissemination Level:
PU Public

Delivery date: **31th December 2017**
Month: **M19**
Organisation name of lead contractor for this deliverable:
Helmholtz Centre for Environmental Research GmbH (UFZ)

This project has received funding from the European Union's H2020 Programme under the topic of potential of current and new measures and technologies to respond to extreme weather and climate events under grant agreement no. 700099.

Document Information

Title	**Report describing the methodology for the co-evaluation of the ANYWHERE distributed platform For Decision support in Emergency Management Operation Services (A4DEMOS)**
Lead Author	Oliver Gebhardt
Contributors	Annemarie Müller Christian Kuhlicke Annakaisa von Lerber Cristina Vicente Moreno Galateia Terti Giovanni Napoli Henny van Lanen Ilona Láng Isabelle Ruin Marc Berenguer Nicola Rebora Nils Hählen Paolo Ciavola Xavi Llort Shinju Park Daniel Sempere-Torres
Distribution	PUBLIC
Document Reference	

Document History:

Date	Revision	Prepared by	Organisation	Approved by
29.09.2017	Rev_1_0	O. Gebhardt	UFZ	C. Kuhlicke
13.10.2017	Rev_1_1	A. Müller	UFZ	O. Gebhardt
27.10.2017	Rev_2_0	O. Gebhardt	UFZ	A. Müller
10.11.2017	Rev_2_1	C. Kuhlicke	UFZ	O. Gebhardt
29.11.2017	Rev_2_2	O. Gebhardt	UFZ	C. Kuhlicke
09.12.2017	Rev_3_0	C. Kuhlicke	UFZ	O. Gebhardt
11.12.2017	Rev_3_1	O. Gebhardt	UFZ	C. Kuhlicke
20.12.2017	Rev_3_2	A. Sanchez, S. Park	UFZ	D. Sempere
21.12.2017	Rev_3_3	O. Gebhardt	UFZ	C. Kuhlicke
21.12.2017	Rev_4	D. Sempere-Torres	CRAHI-UPC	S. Park

Related Documents:

This report and others are available from the **ANYWHERE** Project Website at: http://www.anywhere-h2020.eu/

Table of Contents

1 Rationale of co-evaluation

1.1 Support of the development process

ANYWHERE strives to organize its innovation and development process in a collaborative manner; therefore, it includes not only researchers throughout the project but also developers, potential users and other stakeholders that might share an interest in the various outputs developed by ANYWHERE. The ultimate goal of ANYWHERE is to empower responder institutions and exposed people to enhance their anticipation and pro-active capacity of response, including citizen's self-preparedness and self-protection, to face extreme and high-impact weather and climate events and thus, to save lives.

Therefore, we will ensure that the outputs developed by ANYWHERE reflect the needs and requirements of potential users and have a high potential for being exploited after the end of the project. In order to provide a tool that is considered useful and usable by users it is of utmost importance to thoroughly include and respond to user's needs, expectations, and expertise when developing the platform. This can be achieved if project partners participate in the project as co-producers of the platform, meaning that they are closely involved in its development process.

Evaluating both the development process and the products of ANYWHERE conjointly with all project members gives highly valuable feedback to all involved parties on how the project with its collaborative character is progressing and what needs to be adapted over time.

1.2 Support of the sustainable implementation process

Consultations with ANYWHERE partners for developing D1.1 and D1.2 revealed that compliance of the project's products with the rather technical requirements and specifications is at best a necessary but not a sufficient precondition for their sustainable implementation at the different pilot sites. Therefore, it is promising, if not required, for ANYWHERE's innovations to actually be used beyond the project's lifetime to enquire the determinants of (successful) innovation diffusion and make use of these knowledge stocks to promote the sustainable implementation of ANYWHERE products.

Everett M. Rogers' "Diffusion of Innovations" (first edition 1962) is the seminal work in this research field. Over the decades some of the basic assumptions of the innovation diffusion model (e.g. linearity of the innovation process) as well as the oversimplified analysis of negotiation processes within organisations and privileged focus on individuals (e.g. Van de Ven 1991) were critically discussed. Still, Rogers' approach to empirically analyse attributes of innovations, which positively or negatively influence their chance for adoption, was widely appreciated.

Rogers considers innovation process as a sequence of:

- "Idea invention"

- “Development”
- “Adoption/diffusion”

According the Roger’s model adoption of an innovation comprises of the following steps:

- “Knowledge/awareness”: User/decision maker becomes aware of the innovation, its functionalities, potential benefits etc.,
- “Interest/persuasion“: User/decision maker forms a first opinion about the innovation,
- “Evaluation/decision“: User/decision maker becomes active, i.e. pursues actions to adopt or reject innovation,
- “(Trial) implementation“: User/decision maker applies innovation on a trial basis,
- “Confirmation/institutionalization“: User/decision maker institutionalizes or stops the implementation of an innovation based on the feedback received, e.g. from outside or inside the organisation.

Figure 1 provides an overview of Roger’s model and the innovation attributes influencing this process.

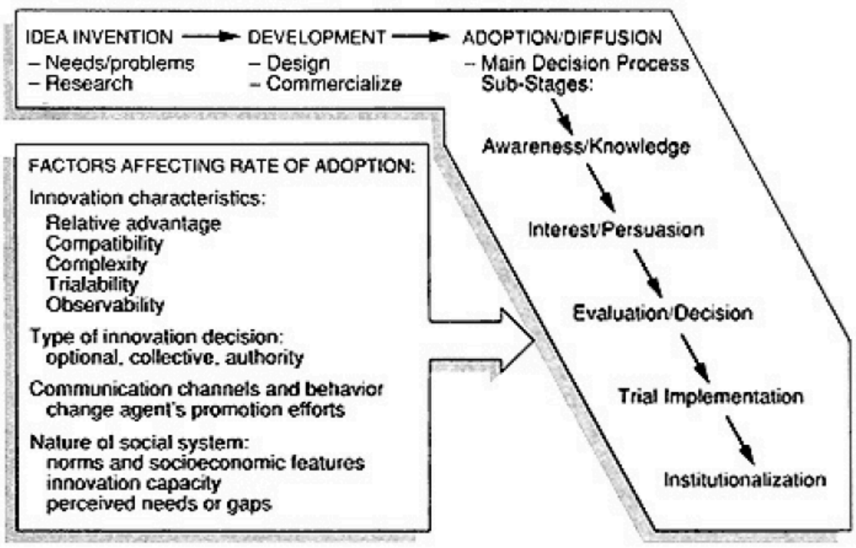

Figure 1 Combination of Rogers's basic stages of the innovation process with factors that predict the rate of innovation adoption (Van de Ven 1991, p. 135).

The model has been tested in many different contexts. On this empirical basis it was concluded that there is strong evidence that the lion's share of variance in the rate of adoption of an innovation and its diffusion can be explained by the inner-organisational perception of five innovation attributes (Rogers 2003):

- Relative advantage, i.e. degree to which an innovation is perceived as better than the idea it supersedes.

- Compatibility, i.e. degree to which an innovation is perceived as consistent with the existing values, past experiences, and needs of potential adopters
- Complexity, i.e. degree to which an innovation is perceived as relatively difficult to understand and to use.
- Trialability, i.e. degree to which an innovation may be experimented with on a limited basis.
- Observability, i.e. degree to which the results of an innovation are visible to others.

Co-production and co-evaluation can enhance the degree of adoption and diffusion of the ANYWHERE platform A4DEMOS. There are various ways, in which the perception of these attributes can be influenced in the development and implementation phase. All five attributes are addressed, whereas the co-evaluation of A4DEMOS primarily focuses on the first innovation attribute, i.e. the perception of the relative advantage of a disaster risk management system including A4DEMOS over the current operational systems without A4DEMOS at the particular pilot sites.

The evaluation of the **relative advantage** of the new system will be based on the performance criteria selected and weighted by ANYWHERE's operational partners, i.e. future A4DEMOS users. This comparative assessment enhances the visibility and, thereby, the perception of relative advantage of A4DEMOS by individual operators as well as by the respective organizational unit (and potential new customers).

The collaborative nature of the development process offers the opportunity to also bring into focus the other four innovation attributes.

In ANYWHERE **compatibility** with the organisational context, i.e. the regulations, values, believes, experiences, and needs, of the future users is supported through thorough context and needs analysis carried out in WP1 and documented in D1.2.

The ability of the users to understand and use the platform, i.e. their perception of its **complexity**, is enhanced through the pilot site training activities during the implementation period. Furthermore, developers maintain close contacts to operational partners also providing support during the demonstration period.

Supported testing and demonstration activities will allow future users to experiment with A4DEMOS at their specific working contexts. Thereby, the **trialability** of the innovation will be ensured.

Many ANYWHERE activities are geared towards making the results of the implementation of A4DEMOS visible to others at the pilot sites and beyond, which will guarantee **observability.**

Beyond these 5 innovation attributes according to Rogers there are additional factors influencing the *rate of adoption*, i.e. the relative speed with which an innovation is adopted by members of a social system (Rogers 1983, p. 240):

- Type of innovation decision,

- Nature of communication channels diffusing the innovation at various stages in the innovation-decision process,
- Nature of the social system, and
- Extent of change agents' efforts in diffusing the innovation.

In the context of the ANYWHERE project these factors are given framework conditions, which cannot and are not to be influenced by the project itself. Information regarding these factors can be collected and analysed in order to investigate why adoption and diffusion processes unfold in a particular way at a specific pilot site, but this is beyond the scope of the evaluation activities in ANYWHERE.

1.3 Support of market outreach of ANYWHERE

The co-evaluation of A4DEMOS will support ANYWHERE's market outreach activities by:

- Providing evidence on the changes the use of the ANYWHERE platform can make in real world disaster risk management operations;
- Facilitating judgements on the transferability of evaluation results to other, similar decision making contexts by providing comprehensive descriptions of the pilot site-specific context conditions of the evaluations;
- Offering a methodological framework, which can be used by ANYWHERE partners operating case studies (WP5) or the fifth EU pilot site (WP6) to analyse the efforts as well as the benefits the ANYWHERE platform can bring to their activities;
- Support ANYWHERE's market development activities (Task 7.2) by offering a methodological framework, which can be used by potential A4DEMOS licensees to analyse the efforts as well as the benefits the ANYWHERE platform can bring to their activities.

2 Evaluation in ANYWHERE

In this section, we would like to position the evaluation work performed in WP1 against the background of the wider ANYWHERE context. In various working packages steps are undertaken aiming at validating and evaluating the outputs prepared by the ANYWHERE consortium. This includes the following steps performed in different WPs:

- Validation of algorithms;
- Verification and validation of the multi-hazard early warning system (MH-EWS);
- Verification and validation of the disaster risk management platform (A4DEMOS);

- Co-evaluation of the co-production process to support iterative platform development process to ensure that A4DEMOS is in line with needs and expectations expressed by ANYWHERE partners;
- Co-evaluation of A4DEMOS at the pilot sites to make this added value visible for partners to promote sustainable implementation of A4DEMOS at the pilot sites.

Table 1 provides an overview of these activities in ANYWHERE.

Table 1: Overview of different types of evaluations conducted in ANYWHERE

Type of evaluation	Task	What	Leading partner	When
Validation of algorithms	T2.8	Robustness and uncertainty of algorithms to assess weather-related event induced impacts	UNIGE & WP2 team	M9-M36
Verification and validation of the MH-EWS	T3.4	Testing and performance evaluation of the MH-EWS system and modules/tools refining	RINA, HYDS	M07-M37
Verification and validation of the A4DEMOS platform	T3.4	Design of the test plan	RINA	M7-M15
	T4.2	Platform Tests and validation	AIRBUS	M18-M38
	T6.2	Training of the operators and responders' personnel on the customized version of the A4DEMOS platform at the pilot sites	CIMA HYDS FMI UNIGE	M21-M24
Co-evaluation of the co-production process	T1.4	Development of co-evaluation framework	UFZ	M4-M19
	T1.5	Application of evaluation framework for collecting feedback from ANYWHERE partners	UFZ	M20, M29, M35
Co-evaluation of A4DEMOS at the pilot sites (user perspective)	T1.4	Development of co-evaluation framework	UFZ,	M4-M19
	T6.7	Feedback of the operators, first responders and stakeholders recommendations	INTC	M25-M39
	T1.5	Application of evaluation framework for 4ADEMOS at the pilot sites	UFZ	M19-M37

2.1 Validation of algorithms[1]

2.1.1 Background of validation of algorithms

Basic information about the uncertainty and robustness of each of the algorithms has been collated through a survey led by UNIGE/METEODAT. The survey included a questionnaire that addressed floods, heat waves and health hazards, heavy rainfall, storm surges, wind, drought, fires and landslides. Questions were about: (i) uncertainties of input and state variables, (ii) parameter uncertainty, (iii) test on very extreme events (outliers), (iv) assessment of False Alarm Rate (FAR), (v) accounting for cascading events, (vi) future climate conditions. It appeared that the total set of algorithms covers a broad variety (Sections 2 and 3) and not all cover all physiographic conditions across Europe. This challenges the definition of common uncertainty and robustness indicators. Simplification is required for homogenization that likely will follow a qualitative approach. The expected outcome is the suitability of algorithms in qualitative terms for the dominant pan-European physiographic settings for current and projected climates. As an outcome of the 1st phase, the preliminary findings of the survey were reported in Deliverable D2.2 (M18) and uncertainty and robustness will be thoroughly analysed in real events during the demonstration period in the different Pilot Sites.

Mapping of the suitability of algorithms in qualitative terms (at least as high, medium, low) for the dominant pan-European physiographic settings (also including, e.g. altitude zone, season) will be done in the 2nd phase. This approach will be developed and validated in the ANYWHERE Pilot Sites (WP6) by comparing impacts derived from forecasted natural hazards with observed impacts (e.g. including validation runs, sensitivity analysis, assessment of forecast quality; see additional tables in section 6 Annex). Parallel to investigating on-going conditions, impacts under a future climate (e.g. RCP8.5 to test robustness) will be explored. If possible, outcome from a Regional Climate Model (RCM) will be used to quantify regional changes in inputs and state variables, and associated extreme events. To extend the results presented in D2.2, the Consortium has proposed to add a new Task 2.8: Robustness and uncertainty of algorithms to assess weather-related event induced impacts specifically focused in the exhaustive analysis of the robustness and uncertainties related to the algorithms included in the Multi-Hazard Early Warning System (MH-EWS). This Task 2.8 has been proposed to run from M9 to M36 to take the maximum profit of the demonstrations (analysis of real events, analysis of false alarms and failures, feedback from operators….) and also to use the data collected from the implementation of the MH-EWS to run off-line calculations on robustness and uncertainty analysis. The findings, including potential strengths and robustness, but also improvements will be documented in the new Deliverable D2.5 (M36).

[1] Input for this section was provided by WP2 partners.

2.1.2 Weather induced hazard Forecasting

Table 2 provides an overview of the validation procedures of the weather induced hazard forecasting algorithms that are used in the Multi Hazard Early Warning System (MH-EWS) developed in ANYWHERE. For clarity it has been separated in the validation aspects of the Meteorological models (used as external input of the MH-EWS), the Hydrological models providing riverine flood forecasting (coming from EFAS) and the algorithms for the different Hazard Forecasting integrated in the MH-EWS.

For the 2nd phase of the validation of the hazard forecasting algorithms (T2.8), the skills of the algorithms included in the MH-EWS will be tested against observed natural hazards. This is a continuation of existing practices, such as for floods (EFAS) and wildfires (GEFF), or newly introduced activities (e.g. pan-European storm surges, heat waves, drought). Focus will be in particular on the Pilot Sites, where the natural hazard will be translated into impacts using knowledge on the site-specific vulnerability.

Table 2: Overview validation processes of meteorological, hydrological and MH-EWS hazard forecasts

Meteorological Forecasting and Nowcasting	At the European Centre for Medium Weather Forecast (ECMWF) the Integrated weather Forecast System (IFS) undergoes continuous verification; Summary published annually (Haiden et al. 2016); Webpages showing the verification scores are regularly updated; Further verification is carried out by forecast users and a brief summary of the feedback from this is also published (Hewson 2017). FMI meteorological forecasts from the HIRLAM and HARMONIE-AROME models (Bengtsson et al. 2017) are used. These forecasts are constantly and automatically verified utilizing numerous well-established validation metrics. Reports, papers and presentations are constantly produced. An operational page of FMI verification is running 24/7 and is available on http://meteor.fmi.fi/verif/. The UPC-CRAHI nowcasting algorithm operating at different scales (from regional to continental) and focused very short-range rainfall forecasts based on radar observations of precipitation. The validation of the algorithm implemented in ANYWHERE has been done through: (i) monitoring the quality of the radar-based Quantitative Precipitation Estimates (QPE), the radar QPEs are adjusted after long-term comparison with SYNOP gauge observations; (ii) offline comparison of real-time nowcasts with the corresponding rainfall observations (typically, from radar QPE), assessing since 2012 the variability of the forecasting skill as a function of time and space (e.g. Berenguer et al. 2005, 2011, Berenguer and Sempere-Torres 2013); and (iii) for selected catchments, assessment of the gain in anticipation in the hydrological response obtained by applying the radar-based nowcasting coupled with a rainfall-runoff model (Berenguer et al. 2005, Corral et al. 2009). RAVAKE (FMI) provides radar-based probabilistic precipitation nowcasts at a regional scale. Validation results are not available for the typical short lead time (1-hour) of RAVAKE nowcasts yet due to the small number of gauges providing verifying observations.
Hydrological forecasting	The hydrological model LISFLOOD, which is part of European Flood Awareness System (EFAS), driven with input from the ECMWF-IFS is used for the MH-EWS hydrological forecasts. Forecast quality is multi-faceted and, consequently, there is the prerequisite for using more than one validation measure to assess the different aspects of forecast quality. Most scores have some strengths, but all of them have also weaknesses (e.g. Jolliffe, Stephenson 2012); in the 2nd phase of the validation of the forecasting algorithms in ANYWHERE, the skills of the weather forecasting and nowcasting, as well as the hydrological forecasting algorithms will be continuously tested against observed weather and hydrological phenomena, which are relevant for the assessment of natural hazards, particularly at Pilot Sites.

Flood forecasting	The flood hazard forecasted with the MH-EWS comes from the output of EFAS. The EFAS forecast performance is continually monitored in terms: (i) counting the number of notification alerts sent out (alerts are counted as hits or false alarms in comparison with observed floods) and (ii) skill scores, such as bias, Nash-Sutcliffe efficiency and continuous ranked probability scores (CRPS). Often the system is evaluated against its own climatology (Pappenberger et al. 2015a) (for performance results see bimonthly bulletin; https://www.efas.eu/efas-bulletins.html; for scientific literature see Alfieri et al. 2014, Pappenberger et al. 2016); new user-focused scores are developed as needed (e.g. Cloke, Pappenberger 2008, Pappenberger et al. 2008); there are inter-annual variations; Further details can be found in Smith et al. (2016).
Flash flood forecasting	FF-EWS module that forecasts flash floods has been validated over significant flood events, through (i) focusing on the ability of the algorithm to identify the affected areas, the timing and the magnitude of the forecasted hazard (Corral et al. 2009; Alfieri et al. 2011, 2017, Versini et al. 2014); the algorithm is currently applied at European scale in real time (Park et al. 2017), (ii) comparing with the EFAS EPIC and ERIC flash flood forecasts (based on probabilistic NWP forecasts), studying the coherence and complementary between the two systems (Alfieri et al. 2011, Versini et al. 2014, Park et al. 2017), and (iii) comparing in a few control points (Corral et al. in preparation).
Landslides & Debris flow	Validation of the landslide and debris flow forecasting has been done in the Catalan Pyrenees (inside the Pilot Site of Catalonia) in 2 ways, i.e. (i) analyzing the performance of the algorithm for the most significant reported events (Berenguer et al. 2015, Hürlimann et al. 2016), and (ii) a systematic evaluation in a monitored catchment, comparing in-situ observations with the forecasted hazard level.
Storm surge forecasting	The validation of the forecasts of the European storm surge model (CFR) is carried out by comparing the simulated storm surge against water level time series available from the JRC database of tidal gauges that consists of more than 200 tide gauges; For multi-annual validation, the ECMWF ERA-Interim (EI) atmospheric forcing is used (mean sea level pressure, wind velocity over the period 1979 to 2016); Since the operational storm surge forecast uses the high resolution atmospheric forecast (ECMWF-HRES), a comparative analysis of the model performance using the HRES and EI will be carried out. The skills of the algorithm will be evaluated in terms of: (i) the root mean square error (RMSE), (ii) relative root mean squared error %RMSE, and (iii) Pearson correlation coefficient (r) for the storm surge. For the tide gauge stations that contain enough data points an extreme value analysis will be carried out; Additionally, an analysis of the alerts counted as hits or false alarms will be done, comparing the algorithm output with the tide gauge measurements. The CFR Regional Storm Surge Model will be validated following the same procedure as for the European Storm Surge model, but using only the tidal gauge station located at the Stavanger harbour (Norway). The validation of the CFR storm surge Local Inundation-Erosion model will be performed based on historic erosion and/or

	flooding events in Stavanger. Historical information of the erosion and/or extension and magnitude of the flooding will be compared with the forecasts.
Heat wave forecasting	Sensitivity plots were constructed to understand the relationship between the weather forecast and heatwaves, i.e. between weather variables used to generate the Universal Thermal Climate Index (UTCI) and the UTCI values themselves (Pappenberger et al. 2015), the FAR of the heat health algorithm will be assessed in the 2nd phase using a 2x2 contingency table, i.e. hits/missed alarms/false alarms/correct rejections, where the heatwave “event” is defined on the basis of a specific UTCI threshold (for example, for strong heat stress UTCI>32 °C).
Wildfire	The wildfire hazard forecasted with the MH-EWS can come from the Global ECMWF Fire Forecasting system (GEFF), which is an extension of the European Forest FIre System algorithms (EFFIS), or from the RISICO algorithm. Any assessments of predictability have to take into account that fire events are rare, thus estimating the quality of predictions is complicated. Using metrics designed for rare events such as the Extremal Dependence Index (EDI) a full assessment of GEFF potential predictability has been performed (Di Giuseppe et al 2016). General findings show that the EFFIS products are useful for both disaster management and planning preventive actions to confine fire-related damage. However, the system’s predictive capabilities are geographically uneven. Where vegetation fuel availability is large, such as in boreal forests, the Mediterranean, South America and Central Africa, fire events are mainly the result of drought conditions and predictability is high. On the other hand, in temperate regions where vegetation fuel availability is limited, such as the mountainous regions of Central Europe, outbreaks of fire can depend on highly variable conditions, such as the short-term superficial drying of the available organic matter on the ground. In these ecosystems the system’s predictive power tends to be lower. The tools developed for the validation of the modelling components of EFFIS are available as open source software under R licence (Vitolo et al. 2017a, b).
Seasonal drought forecasting	Knowledge on the skills of multi-monthly and seasonal drought forecasting is still rather limited; forecasts are highly uncertain (Vitart 2014); Some encouraging initiatives at the national scale are ongoing, as reported by the Hydrological Ensemble Prediction Experiment (HEPEX) community (e.g. Prudhomme 2015, Meißner et al. 2015). Lavaysse et al. (2015) have shown that 40% of meteorological droughts can be correctly forecasted one month in advance over Europe; Arnal et al. (2017) use, similarly to ANYWHERE, the EFAS streamflow forecast driven by ECMWF to investigate forecast skills. Recently EDgE (funded by ECMWF’s Copernicus Climate Change Service) has released new pan-EU hydrological seasonal forecast data for the water sector, including skill and uncertainty information, and a powerful Map Viewer; Testing of the seasonal drought forecasting in the ANYWHERE Pilot Sites will contribute to knowledge on how skilful these are (Sutanto et al. 2017).

Convection storm nowcasting	The convective storm nowcasting method produces yes/no forecasts and probabilities for a storm cell being above a given location and estimate of the severity. The nowcasts were verified firstly by deterministic approach by extrapolating the motion of the storm cell ellipses using their velocity estimates without adding random perturbations. The critical success index (CSI) was used for verifying the deterministic nowcasts. There were no significant differences between different areas in Europe, except the low CSI scores in Central Europe. It appeared that nowcasts of convective storms are generally usable up to 30 minutes, but their reliability drops rapidly after that. These findings were based on radar data (OPERA). When no or no reliable radar data is available lightning densities data were used for the validation of the nowcasts of convective storms. Although, the rainfall field computed from the lightning density cannot describe the fine-structure of precipitation, these still provide areas influenced by intense rainfall outside radar coverage, as well as peak intensities of rainfall.
Gust forecast	The FMI deterministic gust algorithm, which has been implemented over Fennoscandia, and has been validated through comparing the forecast to in-situ observations and other gust forecasts (e.g. ECMWF, HIRLAM, MEPS).
Snow and ice load	Validation of the FMI deterministic snow and ice load forecast, which is based upon radar observations and FMI´s operative weather forecast, has been performed over Finland. It has been done through: (i) algorithm verification (Hoppula 2005), and (ii) monitoring the quality of the radar-based precipitation estimate (Saltikoff et al. 2015).
Precipitation type forecast	Validation of the FMI deterministic and probabilistic precipitation type estimates, which are based upon a number of Numerical Weather Prediction (NWP) models (ECMWF, HIRLAM and GFS), has been done for the winter season 2015-2016. (Nuottakari et al. 2017).

2.2 Verification and validation of the MH-EWS[2]

The goal of the verification and validation activity of the MH-EWS is to check that the system meets the intended specifications and that it fulfils its purpose. Therefore, the verification activity is focused on:

- Provision of a framework to integrate both the forecast and impact models developed within WP2 and the existing Pan-European platforms, complying with the guidelines provided by WP1;
- Provision of the framework to test and validate the MH-EWS (itself and models/tools) in the Pilot Sites (WP3, T3.4)

Figure 2 provides an overview of the timing of the planned verification and validation activities of the MH-EWS.

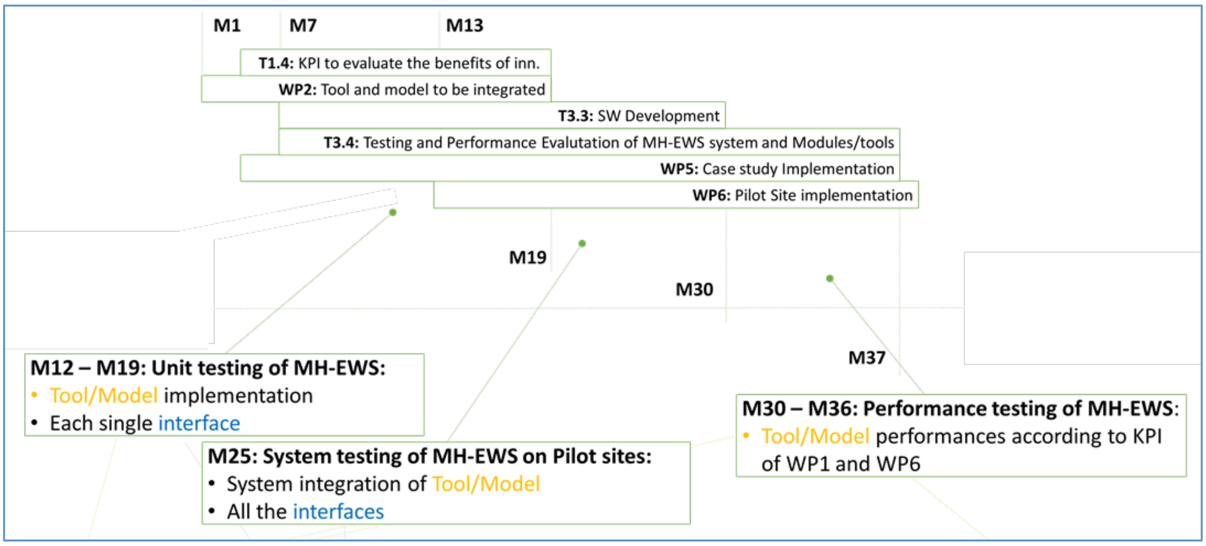

Figure 2: Overview verification and validation of the MH-EWS

First testing phase: The testing activity will be aimed at revealing the presence of software errors and defects considering the system as standalone framework.

Second testing phase: This testing activity checks the proper functioning of the MH-EWS as a whole system at the pilot sites. It will validate the proper interfaces and products provision to the A4DEMOS and self-preparedness tools.

Third testing phase: This activity deals with MH-EWS forecast and impacts performance assessment. It involves more than one partner and it assesses the compliance of the system with MH-EWS requirements and specifications with particular focus on impacts prediction capabilities. Such capabilities are related to WP1 requirement and pilot site experience.

[2] This section was provided by RINA, i.e. the ANYWHERE partner responsible for verification and validation of the MH-EWS.

2.3 Verification and validation of the A4DEMOS platform[3]

Examining a system to see if it does not do what it is supposed to do is only half the battle; the other half is seeing whether the system does what it is not supposed to do.

The A4DEMOS Platform will be validated according to the use cases identified in the site survey carried out directly on the pilot site (WP4, T4.2). These experiences allow us to understand the users' needs and how the knowledge in the ANYWHERE consortium can match with current practices of crisis management.

The main validation activities for A4DEMOS will be based directly on the pilot site experience. This feedback will be based on the user perception and qualitative validation:

- Is the presentation of the A4DEMOS platform comprehensible?
- Are the outputs acceptable for end users?
- Do you have any suggestion to refine the tool?
- Does the tool consider effectively most of the data available that could improve the service output?
- User interface and usability
- Will the A4DEMOS tool replace another tool currently used during crisis management?

To gather these feedbacks already during the project development allows us to understand the efficacy and correct the tuning phase with respect to the control room real needs.

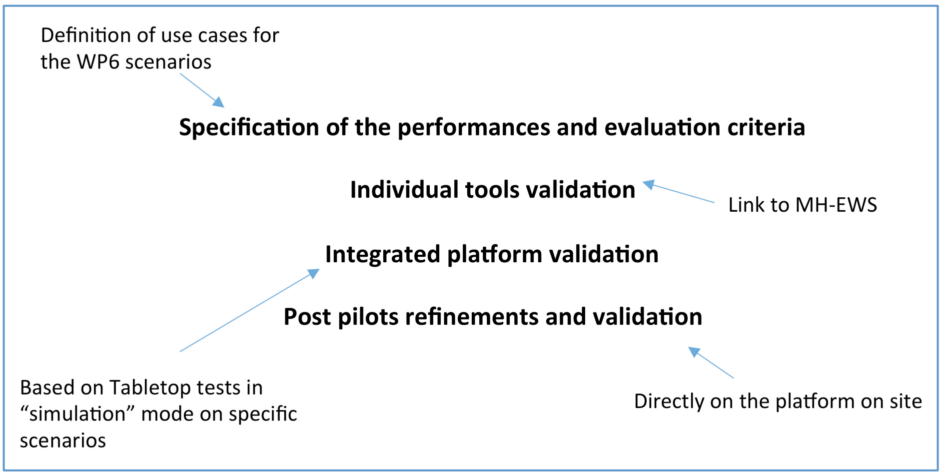

Figure 3: Overview validation of A4DEMOS

The pilot site technical reference within the project will be involved in this feedback collection (T6.7).

[3] This section was provided by RINA, i.e. the ANYWHERE partner responsible for verification and validation of the A4DEMOS platform.

2.4 Co-evaluation of the co-production process

During the first project months all project partners were asked to express what they expect from the project with respect to inputs they are able to provide, the co-production process itself as well as with regard to the expected outcomes. In a second, more elaborated step, users at the four pilot sites were asked to detail those needs that would help to improve the forecasting and emergency management procedures at their sites. Therefore, workshops were organised at all pilot sites.

As a result, a broad overview about needs and expectations by the project partners was compiled and synthesized in D1.1 (needs mainly related to the process, partly also to inputs and outputs) and a short list of main needs was detailed for each of the four pilot sites in D1.2 (needs related to the final product).

In practice, co-production means that those needs, expectations, and also the expertise of the partners are reflected in the platform and that information related to the anticipated content and functionalities of the platform and about the development progress of its development is shared among all parties. The aim of this part of the evaluation is to assess whether the co-production process is successful and whether partners involved in this process have the perception that their needs are understood and considered by those partners responsible for developing the A4DEMOS platform. That includes, more specifically, evaluating to what extends needs and expectations with regard to inputs and co-production processes are met, e.g., the level of interaction, involvement, trust, reflection, and openness shaped the co-production process. This evaluation is performed annually based on a questionnaire (see Table 7, in Annex).

2.5 Co-evaluation of A4DEMOS

The conjoint evaluation of A4DEMOS, i.e. the result of the co-production process, aims at investigating the platform's added value for users through comparative evaluation of their current operational systems and site-specific set-ups of respective disaster risk management system including A4DEMOS functionalities.

The evaluation builds upon the understanding the current working routines, settings and demands for an improved disaster risk management system. The operation of the current management system as well as the system including A4DEMOS are conceptualized and described in detail as baseline and ex post scenario, suitable and feasible indicators to measure the systems' performances are selected by pilot site partners. An adequate evaluation method for the comparative assessment is chosen and data on the performance of the systems for each criterion as well as information on the users and/or decision makers' preferences is collected. Finally data on single performance aspects is aggregated and analysed to determine not only the overall best performing disaster risk management systems, but also to highlight performance differences for specific impact areas or particular user and/or decision maker groups.

3 Framework for the co-evaluation of the co-production process

This section summarises the methodology and framework to be applied for the co-evaluation of the co-production process within the project. It is split in two parts: the co-evaluation of how needs and expectations related to the development process were considered in the implementation process of the platform and the co-evaluation of how the specific needs at each pilot site are reflected in ANYWHERE's product A4DEMOS.

3.1 Co-evaluation of needs and expectations related to the development process

As a theoretical backbone, the co-evaluation of the needs and expectations related to the development process addresses the columns of ANYWHERE's collaborative framework, which was elaborated based on desktop research and a very first survey among the project partners conducted in September 2016 (see D1.1).

3.1.1 Data base

The results of this first survey showed that expectations concerning the development and innovation process of the project are very high: Most quotes addressed the internal governance, i.e. the co-working within the consortium.

To formalise this co-working process a collaborative framework was established with the aim to provide guidelines for a successful co-production. The framework is composed of five cornerstones:

- Establishing and refreshing **trustful relationships** both among project partners but also between project partners and stakeholders as well as among the latter;
- Establishing a transparent **baseline scenario** with regard to project partners' expertise, roles and expectations and with regard to (external) stakeholders' expectations concerning their degree of involvement and the expected outputs developed by ANYWHERE; regularly update the baseline scenario as the project is progressing;
- Establishing regular possibilities and venues for **interaction and involvement** both within larger groups as well as within smaller and more informal and product and tool-oriented settings; this includes ideally face-to-face meetings but may also comprise other forms of interaction (e.g. phone call, webinars etc.) and take place in different degrees of intensity;
- Reserving time for **reflection and open discussion** on how the project is progressing and whether the collaboration process needs to be adapted.

The fifth cornerstone "Selecting and including relevant **project partners and** additional **stakeholders** that represent diverse professional backgrounds" is not considered for the evaluation as it is not subject to change.

3.1.2 Co-evaluation methodology, data collection and analysis

Based on the cornerstones of the collaborative framework, also reflecting the results from the initial 2016 survey, the pilot site visits, and on an informal feedback round on the co-production process during the second project workshop (Helsinki, 19-22 September 2017), a structured questionnaire was prepared (see Table 7 in Annex). This questionnaire serves (a) to describe a new baseline and (b) as a tool to annually evaluate the co-production process within ANYWHERE.

Thus, the reference and (new) baseline for the co-evaluation of the co-production process are the results of the survey, which will be conducted in January 2018. The questionnaire was pre-tested by six members of the consortium representing different groups (developer, pilot site partner/future user) and finalised based on their feedback.

If not formulated differently in the question, the respective partner is asked for his or her individual perception and opinion. For that reason, every person rather than just the different institutions in the consortium is invited to complete the questionnaire. Most of the questions are closed offering predetermined options to answer. However, each section also provides the opportunity to add comments.

In terms of content, the partner is first asked to specify his or her role in the project and association to WPs. The next section of questions aims at finding out if partners are aware of the objectives of the project and if they think that these can be met. In the next section the focus is on the level of involvement, i.e. the personal role and individual contributions to the project. The next set of questions addresses communication and information flow, the last section collects feedback for the pilot site visits and can be skipped by those partners that did not participate in those visits.

This survey will now be repeated annually within the consortium so that the development of the co-production process can be tracked and comparatively analysed over time. The answers are treated anonymously.

The questionnaire is available online and an invitation to complete the questionnaire will be sent to all members of the consortium on a regular basis.

The data will be analysed using SPSS. The results of the survey itself, but also the changes over time will provide valuable insights. The focus of the analysis will be on presenting the distribution of answers for each question graphically in order to highlight which aspects of the co-production process are working well and which ones still need improvement. The consecutive surveys will always include an analysis of how the answers have changed compared to the previous years.

The results of the online questionnaire are summarised in a feedback report, which will be provided to the consortium annually and can be used as tool to further adapt the course of the project, if needed.

4 Framework for the co-evaluation of the ANYWHERE MH-EWS and A4DEMOS

4.1 Foundation of the ANYWHERE evaluation methodology

The collaborative evaluation of the ANYWHERE Platforms, including MH-EWS and the different versions of A4DEMOS adapted at any of the pilot sites, constitutes the second pillar of the co-evaluation activities in the AYNWHERE project. While the first pillar relates to the collaborative process itself, i.e. how these platforms are developed, the co-evaluation of the platforms focuses on potential or actual benefits that result from the use of the platform versions and of the products served by the MH-EWS. Therefore, we predominantly focus on the users of the platform and how they perceive the added value on the individual as well as on the organisational level.

This focus is also grounded in insights from studies trying to evaluate the economic benefits of warning systems at the community or regional scale. These suggest that it is hardly possible to determine the damage reducing effects of the local flood warning system with a conventional damage evaluation approach as the attribution of effects to the warning system is characterised by high levels of uncertainty. Therefore, evaluation methods were developed specifically for warning systems. Pioneering work goes back studies conducted by Penning-Rowsell and colleagues (1978) as well as by Parker (1991). They estimated the damage reducing effects to domestic residential inventories, which was updated in a later model based on a large-scale household surveys. In this study (Parker et al. 2007), respondents were asked questions through a survey about their flood-warning related behaviour. They were presented with a list of about 100 movable objects and were asked which assets they moved in a past flood event in order to locate assets from a potentially flood exposed location to a safer location after they had received a warning. The resulting damage savings were monetised using standard asset values as used in the Middlesex University, Flood Hazard Research Centre flood damage database. However, even such an elaborated and time-consuming approach has clear limitation. Firstly, it focuses exclusively on the damages avoided by moving or raising assets and not on damages avoided by water introducing buildings (e.g. through sandbagging, pumping etc.) or by mobile walls or barriers, which are built-up after a warning was published. Secondly, the study reveals quite clearly that behaviour in response to flooding and the damage savings therefore made are complex matters, and most of the variability in individual responses remains unexplained by the model underlining that even highly-elaborated, survey based evaluation methods are hardly able to explain individual behaviour with regard to flood warning and, thus, the wider regional damage-reducing effects of such warning systems.

Therefore, it appears appropriate to focus on the direct access to the users and the particular relevance of their perception of an innovation and how it is perceived within their organisations to ensure sustainable implementation (see section 1.2) suggests paying special attention to this level of analysis when evaluating ANYWHERE platforms.

The assessment of the changes triggered by research project and their related outputs is a concern of EU Commission for about two decades[4] culminating in the Impact Assessment Guidelines of the EU Commission (EC 2009). But the assessment foreseen for ANYWHERE is not only inspired by these guidelines but also by the profound theoretical groundwork and experiences of various EU projects, which dealt with this challenge in varying contexts. Furthermore, the wider scientific literature engaged with different evaluation procedure, both in the more specific field of early warning but also in the wider field of natural hazards, climate change and environmental research was consulted. In addition, for focussing on the more technological effects international norms, which define software quality[5], were taken into consideration. Finally, bottom-up information collected from ANYWHERE partners in the context of various needs analysis-related consultations were incorporated.

When reviewing the different methods applied the so-called SEQUOIA methodology stands out as a seminal approach. It was originally developed for project evaluation in the context of the EU FP7 research project "Socio-Economic Impact Assessment for Research Projects (SEQUOIA). While the focus of the SEQUOIA project was on the assessment of the potential socio-economic impacts of projects in the area of "Software-as-a-Service and "Internet-of-Services", its application has considerably expanded and meanwhile has been applied several times and adapted by European research projects for assessments in different areas, e.g. for a Pan-European disaster inventory (SecInCoRe), for Digital Cultural Heritage projects (Maxiculture), for Digital Social Innovation projects (iA4Si), for e-Infrastructure projects (ERINA+) and for automated service-oriented architecture testing infrastructures (MIDAS).

The SEQUOIA methodology (Monacciani et al., 2011, p. 4f, Cucco et al., 2016, p. 21) was adapted for the evaluations in the ANYWHERE project. The steps of the ANYWHERE evaluation framework follow the SEQUOIA-logic, but were subdivided and denominations changed to enhance applicability and comprehension. Table 3 provides an overview of the adaptations.

Table 3: Overview steps SEQUOIA and ANYWHERE methodology

SEQUOIA	ANYWHERE
Step 1: Mapping areas of impact	Step 1: Context analysis
Step 2: Baseline identification	Step 2: Description of baseline and ex post scenario
Step 3: Impact indicators	Step 3: Selection of indicators
	Step 4: Selection of evaluation method
	Step 5: Data collection
Step 4: Final analysis, RORI Assessment	Step 6: Data aggregation and analysis

Step 1 includes understanding the current working routines and settings as well as expectations, demands, needs and requirements for an improved disaster risk

[4] For details see Arnold et al. (2005).

[5] This primarily relates to ISO/IEC 9126, which forms the basis of the current ISO/IEC 25000 of the International Organization for Standardization.

management system. It is analysed in which areas and to what extent the current disaster risk management system and the system including A4DEMOS create effects for which stakeholders.

Step 2 encompasses a delineation of the operation of the respective current disaster risk management system (baseline scenario) and the system including A4DEMOS (ex post scenario).

Step 3 contains the selection of suitable and feasible (see section 4.4), indicators to measure the performance of the baseline scenario and the ex post scenario. It is based on the expected effects identified at step 1.

Step 4 comprises the selection of an adequate evaluation method for the comparative assessment of the performance of the baseline and the ex post scenario. The appropriateness of a particular method will be determined by balancing its advantages and disadvantages in the light of the context conditions of the assessments (see section 4.5).

Step 5 contains the collection of performance data and information on the preferences of relevant users and/or decision makers regarding the performance indicators. At this step site-specific sets of indicators determined at step 3 are used to measure the performances of the current disaster risk management system, i.e. the baseline scenario, as well as the disaster risk management system including A4DEMOS, i.e. the ex post scenario. Furthermore, information about stakeholder preferences regarding the indicators and/or criteria applied, e.g. their relative importance, is collected.

Step 6 encompasses the aggregation and analysis of data using the evaluation method(s) selected at step 4 in order to compare the baseline and the ex post scenario. This includes an aggregated overall view as well as more differentiated results addressing performance differences for specific impact areas or particular stakeholder groups.

Figure 4 provides an overview of the consecutive steps of the evaluation.

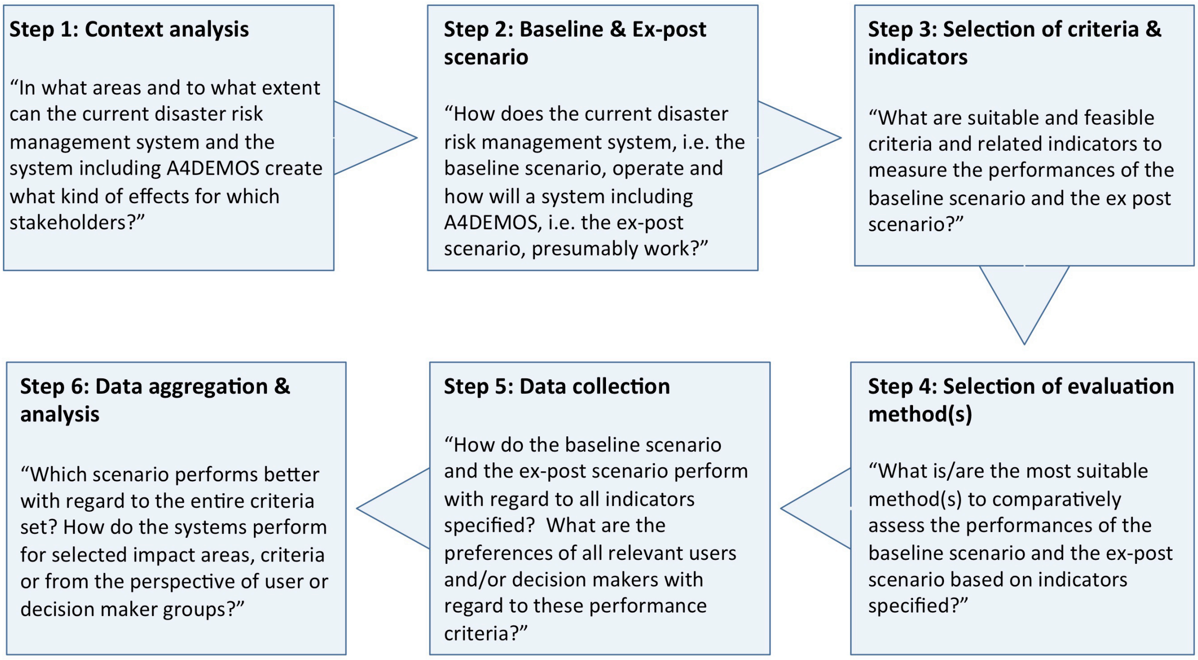

Figure 4: Evaluation flow chart

4.2 Step 1: Context-analysis

The foundation for the context analysis was formed by the comprehensive investigation of the working and decision-making context at the pilot sites during the first partner consultations and field visits. The expectations of project partners as well as their process-related, but also some output-related needs, were investigated and synthesized in section 3 of D1.1: Report with the recommendations and feedback obtained from the Workshop 1. Product-related needs were then focussed in more details in D1.2: Report on needs and requirements from the users (see especially section 6), which also describes in detail the interactions and information flows during the management of hydro-meteorological hazards. Therefore, these two reports form basis for the context-analysis as well the description of the baseline scenario in step 2.

4.3 Step 2: Description of baseline and ex post scenario

The description of the general aspects of the baseline scenarios, i.e. the disaster risk management process at the respective pilot site using the current system, will be based on the data collected for the D1.2. Lessons learnt from the field visits include information on

- Temporal structure of the warning phases and decision-making processes
- Roles, responsibilities of the pilot site partners in the management of hydro-meteorological hazards
- Data sources and information flow during the event

For each pilot site, the training activity (D6.2 and T6.2) provides an overview of the models and data stocks currently used (previous to the implementation of A4DEMOS). A detailed description of the management process for particular event using the current operational system will be produced incorporating the information on the preparation of the re-simulation activities at the ANYWHERE pilot sites.

For the description of the ex post scenarios, i.e. the site-specific set-up of the disaster risk management system including A4DEMOS functionalities, information will be procured from the following sources:

- D1.2: Report on needs and requirements from the users: Description of gaps and key needs (sections 5.4 & 6)
- D5.2: Report describing the products and services developed for the 4 self-protection case studies
- D6.1: Detailed planning of A4DEMOS implementation specifications in each pilot site
- D6.2: Detailed planning of A4DEMOS training of local operators and first responders in each pilot site

4.4 Step 3: Performance criteria and indicators

For the identification of relevant impact areas, evaluation criteria and suitable indicators, which enable us to measure changes of single performance aspects, we again follow the logic of the SEQUOIA methodology, which was originally developed to assess the performance of research projects. But as the focus of the evaluation activities in ANYWHERE is only partly on the project itself but primarily on unveiling changes on the organisational level, i.e. the manifolds effects A4DEMOS has on the activities of the operational users at the ANYWHERE pilot sites. Therefore, an adaptation of the classification of the so-called impact areas as well as the operationalization of the assessment criteria used by SEQUOIA is required to cater for these particular context conditions of the evaluations in ANYWHERE.

At the general level the SEQUOIA methodology differentiates between economic and social impacts and then further subdivides these two main impact areas into the following categories.

- <u>Economic impact:</u> Financial impact, technological impact, environmental impact
- <u>Social impact:</u> Impact on employment and working routine, impact on knowledge production and sharing, impact on social capital

The separate analysis and subsequent division of the *social impact* into these three categories is comprehensible. The categorisation of financial, technological and environmental impact as *economic impacts* appears less convincing as the later two subcategories rather represent distinct categories in their own right. Therefore, we modify the categorisation of the main impact areas and sub-impacts, which we denote criteria, in order to allow as a more nuanced and specific mode of analysis.

Table 4 provides an overview of the commonalities and differences of the SEQUOIA and the ANYWHERE methodology in this regard.

Table 4: Overview impact areas and sub-impacts/criteria of the SEQUOIA and the ANYWHERE methodology

Impact areas & sub-impacts: SEQUOIA methodology		Impact areas & criteria: ANYWHERE methodology	
Economic impact:		*Economic impact*	
	Financial impact		Financial effects
	Technological impact		Economic effects
	Environmental impact	*Technological impact*	
Social impact			Functionality
	Impact on employment and working routine		Efficiency
	Impact on knowledge production and sharing		Maintainability
	Impact on social capital		Usability
			Reliability
			Portability
		Environmental impact	
			Resource use
			System outputs
		Social impact	
			Effects on working routines and employment
			Effects on knowledge production, use and sharing

Impact areas and criteria are defined in the following way:

Economic impact comprises of effects that the systems have on the financial means of the organisations operating them as well as the ones affecting the local, regional or national economy.

- Financial effects, i.e. financial effects of the systems on the organisational level.
- Economic effects, i.e. wider economic effects of the operation of the systems on the local, regional or national level.

Technological impact encompasses various properties of the disaster risk management systems, which characterise different performance aspects. This covers: [6]

- Functionality, i.e. degree of usefulness of the system for its purpose.
- Efficiency, i.e. amount of system resources used for providing required functionality.
- Maintainability, i.e. ability to identify and fix a fault within a system component.

[6]Definitions are partly based on Fleming (2017).

- Usability, i.e. ease of use of the entire system or one of its specific functions.
- Reliability, i.e. capability of the system to ensure service provision under defined framework conditions.
- Portability, i.e. ability of the system to adapt to changes in its environment or of the requirements.

Environmental impact includes effects that the operation of the systems has on the environment and citizens' quality of life through the usage of resources and turnout of by- and waste products.

- Resource usage, i.e. all resources depleted through the operation of the systems.
- System outputs. i.e. all types by- and waste products caused by the operation of the systems.

Social impact addresses effects that the systems have on individuals as well as social groups at different levels of interaction with a privileged but not exclusive focus on the user organisations.

- Effects on working routines and employment, i.e. efficiency of and satisfaction with the disaster risk management process using the systems as well as organisational employment levels.
- Effects on knowledge production, use and sharing, i.e. matters of availability and quality of data inputs for decision-making processes, convenience of information processing, usefulness of system outputs and knowledge exchange capabilities.

These impact areas and criteria are further specified and suitable indicators and metrics are suggested to evaluate the magnitude of change due to the application of A4DEMOS. The list of indicators compiled is based on literature research also consulting publications on the *Indicators of Progress* of the *Hyogo Framework of Action.* A comprehensive overview of the performance criteria and indicators proposed is provided in the Annex (see Table 8).

The indicator list for technological effects is based on ISO/IEC 9126-2. Differences to the current norm ISO/IEC 25010 are limited (see Table 9). Still, after the selection of suitable and feasible indicators by pilot site partners terminologies will be updated to accommodate the current standard.

4.4.1 Selection of performance criteria and indicators

Consultations with future A4DEMOS operators as well as the field visits revealed that the evaluation contexts at the ANYWHERE pilot sites vary substantially due to differences in:

- Disaster risk management procedures;
- Site-specific needs and, therefore, A4DEMOS set-ups;

- Possibilities for running re-simulations of the management of past events;
- Data availability and/or options for collecting performance data;
- User preferences regarding performance criteria.

Many performance criteria and indicators are proposed in the literature or used in research projects, some of which seem to be particularly suitable for assessing the performance of a system against a benchmark or for comparing distinct and clearly defined systems. For the specific evaluation setting in ANYWHERE, i.e. a comparison of two disaster risk management systems of which one is an amended version of the other, some of the criteria and indicators appear to be less suitable.

Still, we compiled a comprehensive repository of criteria and indicators (see Table 8 in Annex) as their selection should be informed by future system users. Given the empirically proven fact the sustainable implementation of an innovative system largely depends on the perception of its relative advantage over the existing one (see section 1.2), their views are of particular relevance. So the selection of measuring instruments is related to their particular needs, which were in the development process translated into specific A4DEMOS functionalities, which in turn have the potential to create an added value for their disaster risk management operations.

The selection of indicators is widely discussed in literature often leading to extensive lists[7] or smaller selections of indicators to be considered named by acronyms such as SMART (specific, measurable, assignable, realistic, time-related – originally project management-focused), RACER (relevant, acceptable, credible, easy to monitor, robust to manipulation, EC 2009, Annex 13, p. 76f) and SPICED (Subjective, Participatory, Interpreted and communicable, Cross-checked and compared, Empowering, Diverse and disaggregated – originally for development context, Roche 1999, p. 49).

Indicators proposed to pilot site partners (see Table 8 in Annex) fulfil the context-independent aspects, e.g. they are specific, measureable, time-related. For considering the particular evaluation context we focus on two selection criteria, namely suitability and feasibility.[8]

Suitability relates to the property of indicators to capture relevant aspects of the added value A4DEMOS components theoretically can provide for the disaster risk management process at the respective pilot site. For determining the suitability the following aspects should be considered:

[7] Many examples for the selection of indicators in the sustainability context, which also could be applied in the disaster risk management context, can be found in Meadows (1998), Bell, Morse (2003) and Guy, Kibert (1998).

[8] For a similar approach see EU FP7 project „Friendly and Affordable Sustainable Urban Districts Retrofitting – FASUDIR“, Deliverable 2.4: DST Key Performance Indicators. There are many examples of feasibility being taken into consideration when selecting indicators, e.g. Hyogo Framework of Action mandates UNISDR to develop a set of “generic, realistic and measurable indicators, keeping in mind available resources” (HFA 2007, p. 17).

- Appropriateness to determined changes (considering improved/additional functionalities)
- Conformity with existing evaluation schemes (e.g. internal performance assessment schemes)
- Relevance for real world decision making

Feasibility primarily addresses the data collection process, i.e. it implies reflection on:

- Data availability
- Opportunities for data collection in the context of ANYWHERE training and testing activities

Pilot site partners will select adequate, i.e. suitable and feasible, criteria for their specific setting by rating the indicators proposed (see Table 8 in Annex). For this exercise a template will be provided (see Table 5).

Despite the use of context-specific criteria and indicator sets we suggest that a core set of criteria will be applied for the evaluations at all pilot sites. Figure 5 provides an illustration of this approach.

Table 5: Template for rating suitability and feasibility of potential key performance indicators

Impact areas / Effects	Criteria	Indicators	Suita-bility (1-4)	Feasi-bility (1-4)	Comment
Economic impact					
Financial effect	Operational costs	Total annual costs occurring at your organisation due to operation of the system.			
…	…	…	…	…	…
Technological impact					
Functionality	Accuracy	Frequency of incorrect or imprecise result caused by operation procedures.			
…	…	…	…	…	…
Environmental impact					
Resource use	Energy consumption	Energy used to operate the system per hour/day/annum			
…	…	…	…	…	…
Social impact					
Effects on knowledge production, use and sharing	Knowledge production: Output	Level of knowledge about suitable management options.			

…	…	…	…	…	…

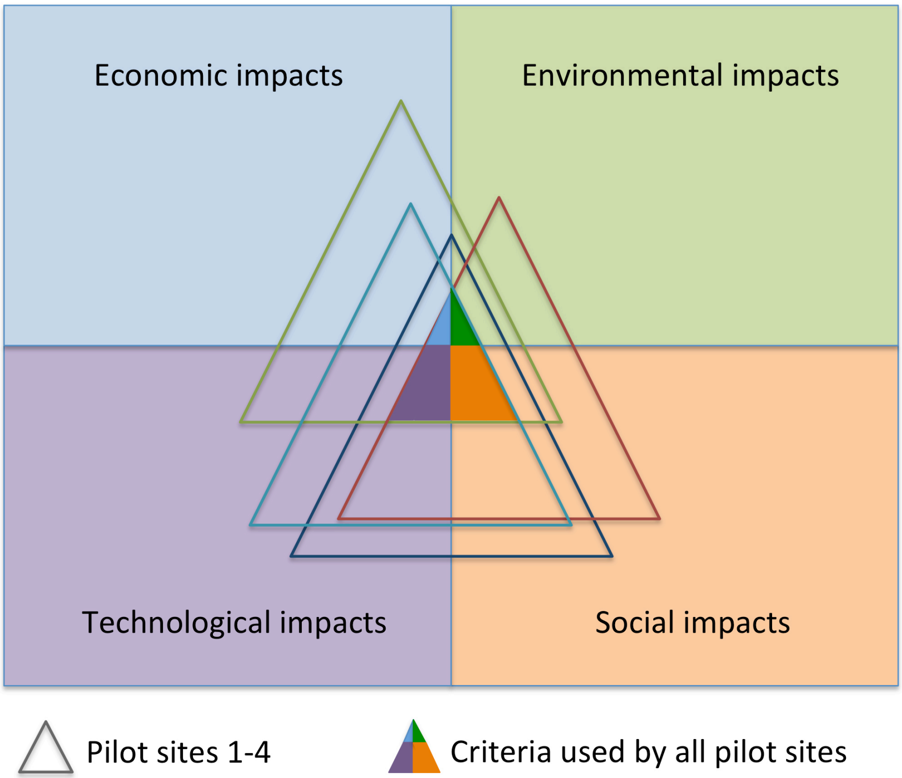

Figure 5: Overview impact areas
Note: Differently coloured rectangles represent totality of criteria and indicators assigned to each of the four main impact areas.

4.5 Step 4 Economic evaluation methods

4.5.1 Cost-benefit analysis

The traditional methodological framework for economic assessment of alternative systems is the Cost-Benefit Analysis (CBA)[9]. Its main objective is to find the most efficient, i.e. optimal option. Its basic rationale is to relate the aggregated benefits of an alternative to its costs to determine its net benefits compared to a reference "baseline" option. If different options are compared the alternative with the highest net benefit is selected.

All costs and benefits have to be included in monetary terms and estimated for each year of the evaluation timeframe. Typically, investment costs occur in year 0, while current costs are accounted for annually, and re-investment costs may accrue after a

[9] For more information on CBA see Hanley, Spash (1993), Hansjürgens (2004), Brouwer, Pearce (2005), Young (2005). For applications in the context of natural hazard management see MAFF (1999), Brouwer, Kind (2005), Pearce, Smale (2005), Turner et al. (2007), Thöni et al. (2009), Meyer et al. (2012, 2014).

certain period of time. Benefits are usually estimated through the reduction of expected annual damage, which is typically determined through modelling. If this is not feasible, benefits can be estimated on the basis of consultations with experienced stakeholders or experts. The annual expected damage value and, thus, the benefits of an alternative are subject to change over time as climate and socio-economic framework conditions change.

In order to make costs and benefits occurring at different points in time during the evaluation period they have to be discounted, i.e. converted into their present value, which in turn is their value at the time the CBA is conducted. The rationale behind discounting is the time preference of decision makers: Benefits as well as cost are valued higher the sooner they are received or have to be paid for, respectively (Hanley, Spash 1993). The choice of a specific discount rate has an important impact on the result of a CBA, i.e. the net (present) benefits determined. Therefore, there is an intensive debate about specifying its level and trend over time (e.g. constant vs. declining).[10]

Typically one of the following two decision rules is applied, when running a CBA.

- Net present value (NPV) test: If the NPV is positive (NPV>0) discounted benefits exceed discounted costs, i.e. the implementation of an option leads to a higher level of social welfare. The option with the highest NPV is selected.
- Benefit-cost ratio (BCR) test: If the ratio of discounted benefits to discounted costs is bigger than 1 (BCR>1) this option leads to a higher level of social welfare. The BCR does not provide evidence on the total size of the social benefit, but just on the relation of the expected benefits to the occurring costs.

Therefore, the result of a CBA depends on the specific benefit and cost curves of the alternatives compared, so that it is possible that options with high BCRs can have small NPVs. It is recommended to use NPV if one option is to be selected and BCR if there is a given budget, which can be used for several options until it is exhausted (Pearce, Smale 2005).

Sensitivity analyses should be conducted to investigate how robust results are to variations in input data, e.g. discount rates.

4.5.2 Cost-effectiveness analysis

Cost-Effectiveness Analysis (CEA[11] is an evaluation approach that relates the (monetary) costs of an alternative to its effects, which are measured using a non-monetary target indicator. So typically the ratio's denominator represents changes of a particular effect resulting from the implementation of an alternative and the numerator represents the costs occurring due to the implementation.

[10] See for example Gowdy (2009), Turner et al. (2007).

[11] For a discussion of CEA see e.g. Messner (2006), Rheinsberger, Weck-Hannemann (2007).

As for the CBA options to be compared for a given evaluation period have to be specified. All expected costs have to be assessed in monetary units and a suitable non-monetary target indicator has to be determined. Such an indicator can be either quantitative, e.g. specific number of people to be protected, or qualitative, e.g. low, medium or high protection levels. As for the CBA discounting has to be applied.

Again depending on the purpose of the evaluation different decision rules can be applied.

- Costs to achieve defined target: Option is selected that achieves defined target value at lowest costs.
- Effectiveness at given cost level: Option is selected that reaches best target value at given costs.
- Cost-effectiveness ratio: Descending selection of non-precluding options (from best to worst cost-effectiveness ratio) until a given budget is exhausted.

Sensitivity analyses should be conducted to investigate results' robustness.

4.5.3 Multi-criteria analysis

Multi-Criteria Analysis (MCA) [12] is an evaluation approach that judges the performance of alternatives against a number of objectives or evaluation criteria (Belton, Stewart 2002). So it can be applied if the assessment of a system either has to take several objectives into consideration or if several criteria covering different aspects of the performance are to be considered. The latter option is particularly relevant if indicators measuring these aspects cannot (easily) be merged into one non-monetary (target) indicator or expressed in monetary terms. So in contrast to the CBA and CEA the availability of monetary or other quantitative data is not an application prerequisite as performance aspects can also be qualitatively assessed, e.g. through expert judgements.

Many different MCA approaches exist to compare pre-defined sets of alternatives. The two most prominent methodological streams are Multi-Attribute Decision-Making (MADM) concepts, which are based on Multi-Attribute Utility Theory (MAUT), and the so-called outranking concepts (see e.g. Keeney, Raiffa 1993, Drechsler 1999, Klauer et al. 2006). In contrast to MAUT-based approaches outranking concepts do not assume that decision makers are completely aware of their preference structure. They perform pairwise comparisons of the alternative options across all evaluation criteria.[13] So only preferences regarding these pairwise comparisons and the relative importance of the various criteria have to be unveiled.

[12] For general description of different MCA approaches see e.g. Bana e Costa (1990), Zimmermann, Gutsche (1991), Munda (1995), Vincke (1992), Belton, Stewart (2002). For applications in the context of natural hazards management see e.g. Bana e Costa et al. (2004), Brouwer, van Ek (2004), Akter, Simonovic (2005), Kenyon (2007), Meyer (2007).

[13] For an overview see Zimmermann, Gutsche (1991)

MCA approaches have in common that they subdivide the overall performance of an option into single performance aspects, i.e. criteria. Performance data for each alternative and each criterion is collected. These criteria values are then normalized and aggregated using the weights, which reflect the importance of each aspect for the overall performance. Criteria weights are to be elicited by those that the evaluation is being done for.

The decision rule to be used to determine the best-performing option depends on the specific multi-criteria approach applied. MAUT-based approaches select the best option on the basis of the highest weighted criteria value sum. Outranking approaches select it using highest outflow, lowest inflow or highest net flow scores.

Sensitivity analyses should be conducted to investigate results' robustness.

4.5.4 Comparative appreciation of evaluation methods and recommendation

Each of the evaluation methods introduced in the preceding section has particular strengths, but there are also some challenges when applying them. This section discusses these characteristics and reflects on the applicability of the methods in the context of ANYWHERE.

CBA[14] is well known and has already been applied in many different contexts, it aims for accounting for all positive effects (benefits) and negative effects (costs) of an option and allows for comparing alternatives based on their net (social) gains. Major challenges are that all relevant costs and benefits have to be expressed in monetary terms, which is very complicated in non-market sectors and for non-technical options. Therefore, often in these sectors only partial CBAs, which then have to be complemented by supplementary evaluation methods covering e.g. distributional issues. The application of a CBA is recommended if (1) the cost perspective is relevant, (2) monetization of all or at least of the most significant performance aspects is feasible and risk probabilities are known and sensitivity is rather small.

The strength of the CEA is that beneficial performance aspects can be considered in non-monetary terms. Still the specification of a single target indicator to measure heterogeneous performance aspects is very challenging. This is particularly relevant if options unfold manifold, complex and/or cross-sectoral direct and indirect effects. In principle the application of a CEA is restricted to comparisons between options that produce directly comparable outputs measured in the same unit (Birch, Gafni 1992). Furthermore, the possibility to consider uncertain data is rather low and - based on its results - no clear judgement of the net (social) gains of the implementation of an option is feasible. Therefore, the use of a CEA is recommended if (1) there is a single objective, which can be operationalized using specific target level and (2) monetary cost estimates are available. Actually its rationale is very similar to the one of the CBA, but it is applied if monetary valuation of the benefit criterion requires disproportional efforts or is simply impossible.

[14] Main arguments are based on Watkiss et al. (2014), European Commission (2009, p. 45f).

The MCA shares the aim to analyse positive and negative effects of options in one single assessment framework to allow for comprehensible comparison of different scenarios. So it also relates to the cost-benefit thinking of the CBA and the CEA but in contrast to those it is more explicitly weighting the pros and cons of options with regard to particular performance aspects. It highlights the trade-offs instead of purely aiming for optimization. Its biggest advantage is the possibility to use monetary, other quantitative and qualitative data for the evaluation. Furthermore, there are some MCA approaches, which are particularly capable of using uncertain input data[15]. They consider this data stochastically using score ranges, triangular distributions or other probability functions. Stochastic PROMETHEE II is an example for such an outranking approach. On the one hand side the possibility to use differently scaled input data for the analysis offers more opportunities with regard to data collection methods, but on the other hand side the elicitation of decision makers' preferences for the (weighted) aggregation of criteria values also requires additional information. The explicit consideration of user, decision maker and/or stakeholder preferences in the assessment process makes MCA a more deliberative and participatory evaluation approach than CBA and CEA. However, for the same reason it is often criticised for being subjective, especially since the methods being used to elicit preferences have not been tested and improved to the extent as e.g. stated preference techniques for monetary valuation. The consideration of time preferences is a challenge if evaluation periods extend far into the future and the analysis of net (social) gains is hampered by the fact that no information on the (net) benefits of the implementation of an option is provided in monetary terms. Still, MCAs aim to provide an overall ranking indicating which of the options performs best considering the preferences of the relevant users and/or decision makers. MCAs provide an additional value as rankings can also be computed for single indicators or specific (groups of) users and/or decision makers, which facilitates the comprehension of and exchange about advantages and disadvantages of one option over another.

Based on the planning of various testing and demonstration activities at pilot site level (training, re-simulation of events), which will provide valuable opportunities to collect performance data for the options "baseline scenario" as well as "A4DEMOS scenario"[16], we expect that the data to be considered for the evaluations in ANYWHERE will be diverse. Different data collection methods will have to be applied and data metrics and scales will have to be dealt with.

Appreciating the pros and cons of all evaluation approaches presented, taking into consideration the lessons learnt from other research projects, having in mind the properties of the performance data (e.g. scales, metrics, level of uncertainty) to be dealt with and reflecting the pre-conditions of the application of the various MCA approaches at hand (e.g. unveiling of decision makers' complete preference structure for MAUT-based approaches), we recommend the use of a probabilistic multi-criteria

[15] Less sophisticated approaches add an uncertainty criterion to the criteria list.

[16] For more information see section 4.3.

outranking approach to comparatively assess the performance of the baseline scenario and the A4DEMOS scenario at the ANYWHERE pilot sites.[17]

4.6 Step 5: Data collection

Once baseline scenarios and ex post scenarios are described, performance criteria and indicators selected and evaluation methods chosen by each ANYWHERE pilot site the data collection process is initiated. Assuming the evaluation will include the application of one of the MCA approaches two types of data, i.e. performance data and data on users and/or decision makers' preferences, are to be collected to assess the performances of the ANYWHERE platform as well as the current operational disaster risk management systems at the different pilot sites.

For every indicator specified to capture a single performance aspect of the baseline scenario and the ex post scenario at each pilot site, information is collected on how the disaster risk management process operates with and without A4DEMOS. For this purpose, selected data collection methods introduced in sections 4.6.2 - 4.6.5[18] will be applied in the context of the training activities and re-simulations of past events planned for the demonstration period at the different pilot sites. Section 4.6.1 is based on the inputs by ANYWHERE partners to D6.2, which shortly describe their planned training activities. As set-ups of A4DEMOS will differ at the each pilot site due to their different requirements each site-specific version of the platform has a particular acronym, e.g. in Catalonia A4CAT, in Switzerland A4ALPS, in South Savo A4FIN and in Genoa A4LIG. A 5th pilot site version for a generic site in Europe will be implemented and tested in the pilot sites of Stavanger (Norway) and North Corsica (France). In addition to the data collection during the demonstration period, performance data will also be collected at the end of the operational testing period, i.e. after a year of operation.

Based on the selection of suitable and feasible indicators performed by the pilot site partners a site-specific strategy for collecting performance data will be specified. Data collection method to be applied, timing and responsibilities are determined for each performance indicator.

[17] This is also in line with experiences from other EU projects such as SEQUOIA and ERINA+, which show that whenever impacts of actions are difficult to monetise and/or only become measurable in the medium or long term MCAs are best suited for such types of evaluation.

[18] Section 4.6.2 introducing the ANYCaRE game was contributed by CNRS and sections 4.6.3 -4.6.5 describing further data collection methods are contributions of UPB. Inputs were revised by UFZ.

Table 6: Exemplary overview data collection strategy

Impact area	Criterion	Sub-criterion	Indicator	Data collec-tion method	Timing	Respon-sibility
Economic impact	Financial effect	Operatio-nal costs	Total annual costs occurring at your organisation due to operation of the system.	Cost account-ting	End of demonstra tion period	Pilot site partners
Techno-logical impact	Usability - Operability	Suitability	Suitability of the system for the specified task(s).	Survey	After re-simulation activity, End of demonstra tion period	UFZ
Environ-mental impact	Resource usage	Energy consump-tion	Energy used to operate the system (hourly/daily/ annually)	Measure ment	End of demonstra tion period	Pilot site partners
Social impact	Effects on working routines and employment	Working routines	Level of satisfaction for execution of compulsory activities under ordinary conditions.	Survey	End of demonstra tion period	UFZ
…	…	…	…	…		…

The investigation and consideration of users and/or decision makers' preferences will be two-fold. Firstly, the suitability and feasibility of performance criteria and indicators are rated using the template provided (see Table 5). This rating forms the basis of the selection of performance indicators. Secondly, the relative weight of each performance indicator and effect category for the overall aggregation is determined by using more sophisticated methods for weight elicitation. These are described in section 4.6.6.

4.6.1 Re-simulations of past events

The objective of the evaluation in the pilot sites is to test the effect of the integration of the A4DEMOS platform in the systems and operations of the different users involved. This will include the monitoring of the performance of the platform in real time during the demonstration period as well as the evaluation of the benefit of the platform on past events compared to the existing systems.

4.6.1.1 Liguria pilot site

The evaluation based on occurred cases will analyse the flash flood event of 9 October 2014. The idea is to reproduce the same event from the forecast of the previous day, then follow the procedural aspect of the Genoa Municipality through the forecasts, observations and actions to be taken during the occurrence of an event (for details for the various steps see Figure 6).

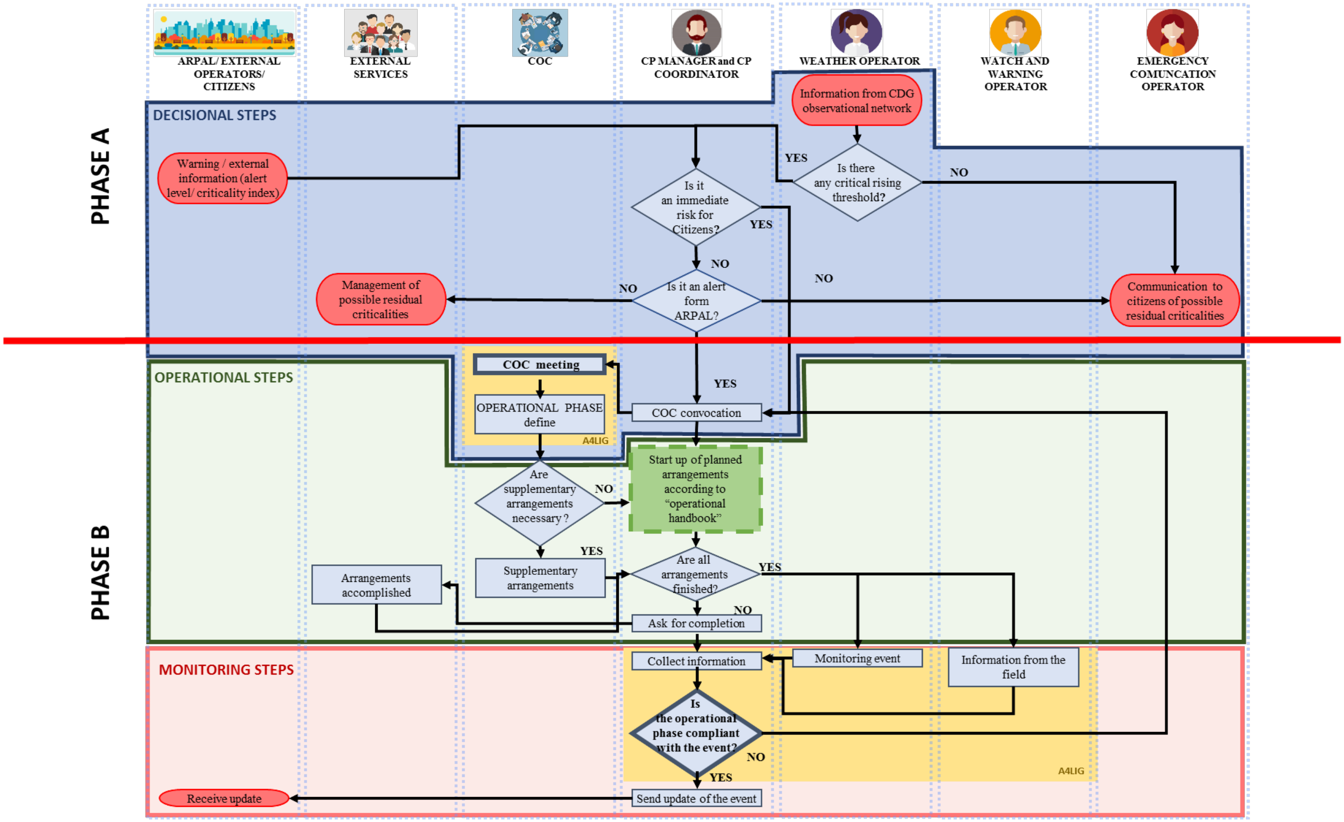

Figure 6: Graphic representation of operational rules

For testing the effects of the A4LIG implementation and changes of the operational activities we reproduce the event in a double configuration:

- Same configuration of the 9th of October 2014
- New configuration with the A4LIG in place.

The event of 9th October 2014 was characterized by heavy rainfall, with very strong intensity, which mainly affected the central part of Liguria and in particular the city of Genoa. The peak was in the evening of October 9th, when an alluvial event affected the capital, and in particular the basin of the Bisagno stream. The event was placed in a highly unstable meteorological setting, characterized by strong stormy activity with stationary structures, though not directly related to the approach or transit of a frontal system.

Precipitation hit the Genoa area in two distinct moments, interspersed with a brief halt of a few hours: a first stormy phase developed between morning and early afternoon; a shorter but more intense phase emerged between the evening and early hours of the following night when rain recorded significant peak times, i.e. among others Genova 141 mm in 1 hour and 226 mm in 3 hours, Geirato 112 mm in 1 hour and 230 mm in 3 hours, Torriglia 88 mm in 1 hour and 212 mm in 3 hours. The event caused one casualty. Overall damages of around 300 million Euros were estimated. Figure 7 provides a chronology of the event.

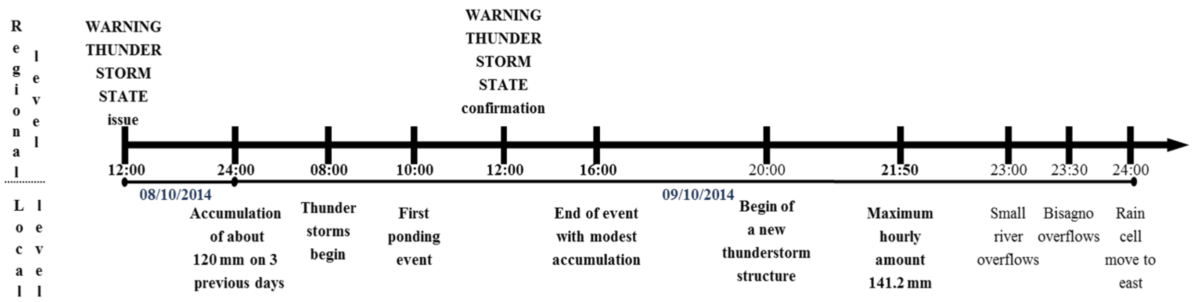

Figure 7: Timeline of the event

4.6.1.2 Catalonian pilot site

At the Catalan pilot site the following events will be re-simulated:

1) Floods (March 2017), i.e. heavy rains, river overflows, waves[19]
2) Wind (December 2014)
3) Forest fires (July 2013), i.e. an urban fire event
4) Snowstorms (March 2010)

In the case of floods, we have chosen a past event (flash flood occurred from 24th to 25th March 2017), to simulate a real situation using two configurations. This procedure will be the same for each hazard and will be based on the current CECAT forecasting system vs. the new configuration with the A4CAT.

The analysis and planning of the emergency actions are also followed with the real-time procedure of the CECAT (emergency management center of Catalonia) through observations (mainly provided from the regional weather services METEOCAT). Relevant steps are displayed in Figure 8.

[19] The event of 24-25 March 2017 is characterized by accumulation situation and rainfall in the central and north coast of Catalonia, which caused significant increases in the rivers managed by the Water Agency in Catalonia - ACA.

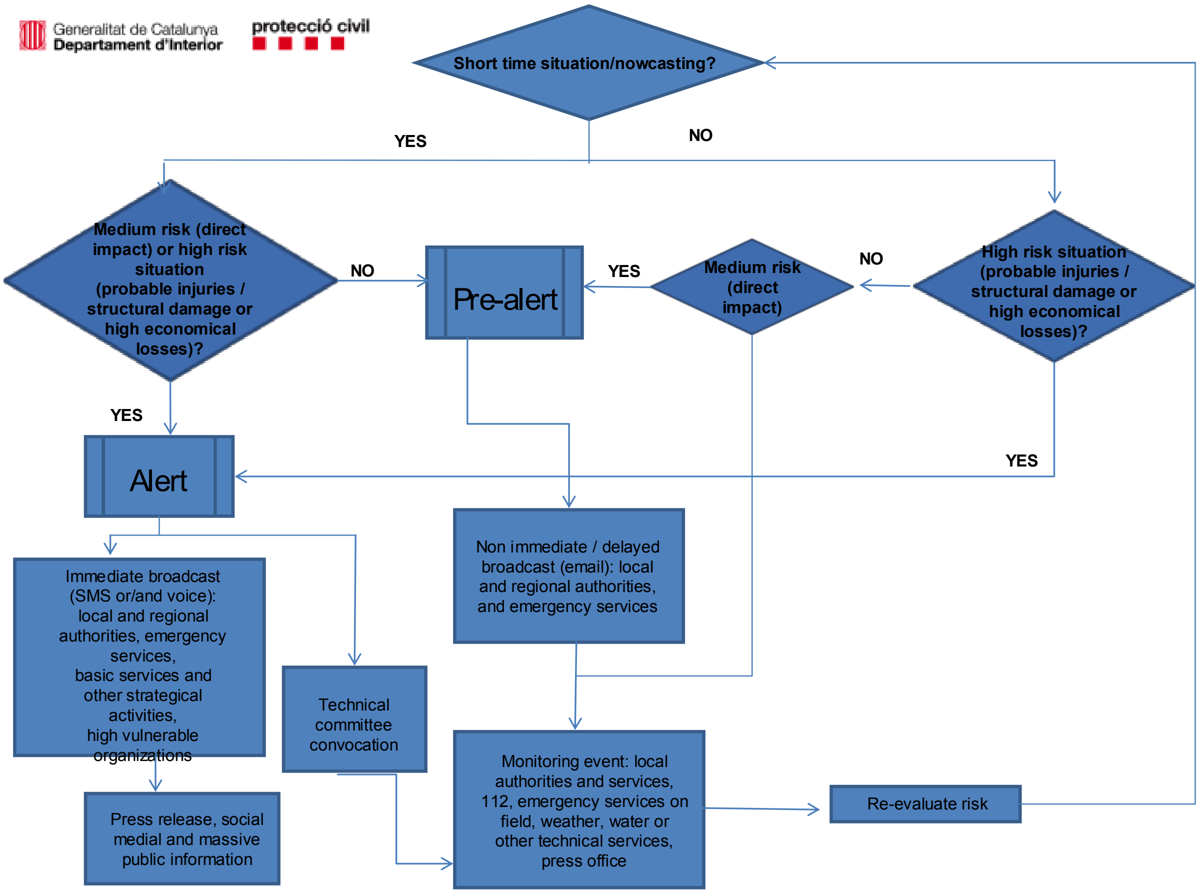

Figure 8: Graphic description of operational rules

By using A4CAT, the aim is to improve the early warning system, before the actual crisis starts, i.e. one hour in advance, before the incidents on the field. We will explore whether A4CAT will allows us to contact municipalities earlier and thus also activate relevant emergency plans earlier. Figure 9 shows the number of calls to 112 related with this episode.

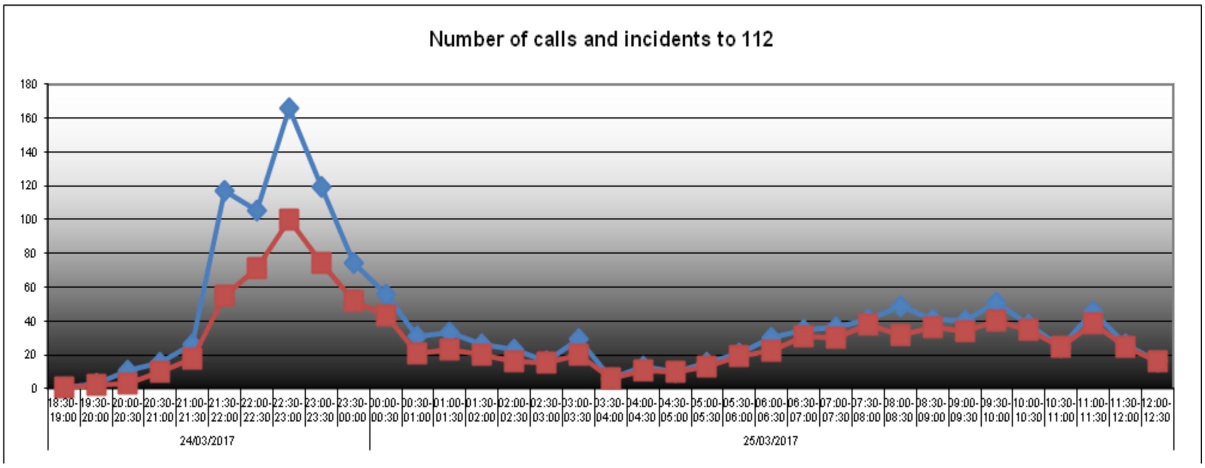

Figure 9: Number of calls to 112 and number of incidents
Note: Number of calls to 112 (in blue). Number of incidents (in red) (from 18:30 LT (UTC+2h) 24 March to 12:30 LT 25 March, 1390 calls to 112 and 1006 incidents)

One indicator will be the time of our response before the incidents on the field will occur. We are also trying to improve through:

1. Integrating all data sources (static and dynamic) that we usually use in these episodes, such as:
   - Static risk information (flooded areas, vulnerable activities, priority points for protection)
   - Weather service forecast
   - Radar (real time)
   - Meteorological (weather service) gauges (real time)
   - River (water agency) gauges (real time)
   - Emergency 112 calls
   - Crowdsourcing
2. Combining real time data sources and translate to potential damage – impact
   - Focus the monitoring on potential critical scenarios (local)
   - Detect and analyse flooded areas
   - Detect critical infrastructures and high vulnerable activities potentially affected
   - Propose warnings and alerts to population
   - In advance: Early warnings
   - With local scale
   - Make it easy: Automatic warnings with only the (summarized) information needed

4.6.1.3 South Savo pilot site

The evaluation activities at the Finnish pilot site are based on historic simulation data. FMI and ISTIKE will use data from two different weather events. The case “Tapani-storm (2011)” is a good example of an intense, large-scale low-pressure windstorm, which usually occur in Finland during the autumn/winter season. In the large-scale weather event (low pressures) the phenomena can be forecasted several days ahead so ISTIKE’s preparation period is rather long. Shorter time-scale actions are simulated with the historical case “Asta-thunderstorm (2010)”. Asta-thunderstorm was an intense convective storm affecting a smaller area than Tapani. Because of the typical nature of convective storms, i.e. rapid development, very short forecast time and intense winds, the impacts of this phenomenon are often severe. Also because the predictability of thunderstorms is so short, the time for ISTIKE to prepare for actions is much shorter than in the large-scale events. Both storm cases are very well documented and provide typical weather-related risks of the Finnish environment in summer and winter situations, i.e. intense rainfall, convective storms, severe winds and heavy snowfall.

The aim of the re-simulations is to show the actual improvement of the end-user operational capabilities before, during and in the aftermath of emergencies induced by extreme and high-impact weather and climate events.

During the pilot phase, i.e. when A4FIN system will be fully operational, ISTIKE will run historical data of Tapani-storm (strong wind) and Asta-thunderstorm through A4FIN. During the simulation we will compare operational procedures, activities of duty officers and activation processes to learn what was done differently during the real cases in 2010 and 2011 at ISTIKE and at the regional Civil Protection authorities' operation centres. If needed, it is fairly easy also to run verification tests (different high-impact weather event) to evaluate benefits of A4FIN.

Ideally, the evaluation should be at a later stage of the pilot phase because then duty officer will have enough experience on using A4FIN. Results of the re-simulations should address differences of speed and reliability of the information received from FMI. This will happen at both ends, i.e. at FMI and ISTIKE, when data is needed to encode only once into the A4FIN. Commanding Civil Protection officers will be able to do faster decisions and create quicker situational awareness.

4.6.1.4 Swiss pilot site

The A4ALPS platform will complement existing tools and, therefore, will only be one of the pillars, experts will base their decisions upon.

The quality of A4ALPS can be described by the degree of uncertainty in decision making with or without the additional information. An evaluation of the reduction of uncertainty will be done during the demonstration phase. This requires that there are one or more events during this phase. The probability of such events is quite high, at least for some of the examined hazards.

4.6.2 ANYCaRE game

ANYWHERE Crisis and Risk Experiment (ANYCaRE) introduces a serious gaming approach to obtain conclusions on "if "and "how" an improved multi-model forecast output, including information on i) impact assessments and maps and ii) live data on exposure and vulnerability derived from social media and crowdsourcing - provided in the frame of ANYWHERE platform - can support the decision chain in European warning systems towards better responses. ANYCaRE is first designed as a tabletop role-playing game[20] (or pen-and-paper role-playing game) for adults, in which participants act their role through speech while sitting in a comfortable setting (Cover, 2005). The game serves as an evaluation tool as well as an interactive communication mean among ANYWHERE's partners and stakeholders.

[20] Pen and paper or tables are not strictly necessary for the game. This term is used to distinguish this format of role-playing game from other formats in which participants act their characters physically as well.

ANYCaRE is designed to simulate one or more of the three main levels in the warning-system decision chain (i.e., 3 groups of roles to be played in the game): i) Level 1: Weather Forecasters; ii) Level 2: Emergency managers/Authorities in charge of civil protection; iii) Level 3: General public and targeted users (private companies). Every group should be preferably played with 10 to 15 players. During the simulation the players are provided with the imaginary case study of "Anywhere City". In ANYCaRE world each game round simulates a successive weekday until the "AnyDay" festival (held on Saturday). Each game day the players receive updated probabilistic forecasts for precipitation and river discharge as well as contextual information for each area in Anywhere City. These data mainly refer to flood early warning products released by the European Flood Awareness System (EFAS) (Smith et al. 2016). As a second trial the players get new ANYWHERE products[21] including improved probabilistic impact-based forecasts, risk assessments and ground observations from social media, and are given the opportunity to rethink and modify their decision if necessary. Each player generates a certain sub-role (e.g., expert hydrologist, school representative, first responder) assigned to him/her by the game moderator, and with his/her co-players has two different issues to address during the simulation. The first is to select one or more of the predefined warning or emergency activities according to the available information and related uncertainties, and the second deals with the communication of those decisions in order to enforce (self-) protective actions.

The first implementation of ANYCaRE was carried out in the frame of ANYWHERE's 2nd workshop in Helsinki (September 2017). A game session was organized with a group of sixteen players to compose the virtual Local Emergency Operation Center of Anywhere City under the threat of (flash) flooding. The aim of the emergency management group was to keep the population safe and ensure smooth execution of everyday life and "AnyDay" festival in Anywhere City, while managing a given budget. Among the players there were PhD students and researchers in weather-related hazards, developers and modellers, emergency managers and operational forecasters. The test experiment was considered as successful since, according to the players, the game 'was very fun' and 'clearly demonstrated the benefit of certain ANYWHERE outputs'. Although still under development, the examples of the ANYWHERE MH-EWS products included in the game were found to reduce the overall uncertainty in the decision-making process. Though, improvement needs were discussed. The players recognized ANYCaRE scenario as very realistic and presented a strong commitment to the storytelling. Participants in Helsinki introduced further aptitudes of ANYCaRE; emphasizing educational scopes such as coaching emergency services in order to sharpen their emergency agility and alertness before the crisis strikes. Applications to other weather-induced risks such as wildfires were also encouraged and are considered for the future. Rather than a single tabletop role-playing, we vision ANYCaRE as a broad experiment campaign that will encompass various versions of gaming (e.g., digital, board games) to be applied at first within ANYWHERE project.

[21] A catalogue of products to be integrated in ANYWHERE Platform is available at http://anywhere-h2020.eu/catalogue/.

4.6.3 Observations

Based on literature research different possibilities to observe users in their environment or during field trails with deviating effects for the observation itself exists. Observation is an important method to gather data about the "real world" and "real needs" of users.

An observation during real emergency incidents is not reasonable in most cases, but there are possibilities to get insight during training exercises.

Main differences are participatory vs. non-participatory observation and direct vs. indirect observation. Hence, before starting an observation it has to be decided how the observer will be involved and whether the participant should be aware of the observation. Documenting the results of an observation can be done using video, voice recording and note taking. Depending on the specific observation protocol types deviate (e.g. diary, notes of observation with context reflection). This includes different methods to collect data from the specific research field.

Aim and applicability of the method

Firstly, it has to be decided to what degree the observer gets involved in a respective situation or training activity. The main challenge of getting strongly involved is to maintain the distance needed to observe and record the on-going situation in an adequate and valid way. In the case of non-participatory observation the researcher does not get directly involved in the action but observes from "outside". Participatory observations imply that researchers collect data during an action under study. (Sekaran 2013). This means observing the activities or situation from "inside" by taking part in the group to be studied. Hence, researchers can better understand the views of participants than an outsider, but at the same time observation objectivity decreases. In non-participatory observations researcher are able to provide a detached and unbiased view of the participants and have time to produce adequate records. In case of a direct, non-participatory observation two different strategies exists:

- Reactive: A reactive observation indicates that participants know that someone is currently observing the situation. In this case the potential of changing the behaviour due to the attendance of an observer need to be taken into account. (Crowther 2005)
- Non-reactive: Non-reactive observation involves serious ethical questions because the study of participants is taken without their awareness. (Bernard 2000)

In addition, indirect observation is a method, which allows researchers to observe outcomes of behaviour rather than observing the behaviour itself. (Bajpai 2011) Sometimes a researcher is unable to observe persons directly, so an indirect observation can be conducted through the analysis of internal organizational documents or other recordings. (Bailey 1994) In this case methods of content analysis are useful.

Application in emergency management

The method observation raises questions on the one hand about the involvement of the observer and on the other hand about participants' awareness of the observation and whether behaviour or routines change due to the observation.

In the context of a real emergency event participatory observation of a first responder is hardly possible, not reasonable and raises ethical, legal and social questions. In some parts of training exercises, which prepare responders for upcoming emergencies, it is possible to be involved. Participation helps to better understand the participant's actions and the complexity of the task to be performed, but hampers a structured recording of the incident flow.[22] For facilitating recording records and documentation of observation, the national research project RescueLab[23] used and developed equipment and methods for semi-automatic recording of exercises.

Strengths and weaknesses of the method

Participatory observation enables the observer to gain insights in participants' actions (complexity of tasks, emotional stress etc.), but the non-participatory observation enhances comparability of observation results and enables the observer to address specific research questions. Use of IT-supported methods to record (e.g. video) can promote a balance between direct, non-participatory and indirect observation (only based on records).

4.6.4 Focus groups

Aim and applicability of the method

Focus group discussions represent a special type of interview. It is a qualitative method for data collection. A small group of people (3 – 12 participants) are focused around a particular topic and a moderator leads the discussion, which typically lasts about 60 to 90 minutes. (CDC 2008, Freitas et al. 1998, Remenyi 2011) Beside this basic version of a focus group discussion a few assorted alternatives exist. The most common ones are: Two-way focus group, dual moderator focus group, duelling moderator focus group, respondent moderator focus group, mini focus group and teleconferences or online focus groups. (CIRT 2017)

A focus group guide presents the basis of the discussion and a facilitator serves the guide as a memory aid and provides a road map. It contains the identification of group members, the type of information that needs to be obtained and how the information collected will be used after the discussion. In the group interview the

[22] The training exercise "Stein der Weisen" 2011 in Dortmund is an example of UPB stuff conducting a non-participatory observation. A large-scale Chemical, Biological, Radiological and Nuclear (CBRN) incident was the basic scenario for the training exercise. Other training exercises, which were observed, include activities at the training centre of the fire department of Dortmund and at the "Akademie für Krisenmanagement, Notfallplanung und Zivilschutz" (AKNZ) served from the Federal Office of Civil Protection and Disaster Assistance (BBK).

[23] UPB was a project partner of RescueLab. The overall aim of the project was to establish innovative education and training methods to support civil first responders.

moderator asks questions about perceptions, opinions and attitudes of participants towards the topic. The method aims at enabling the researcher to obtain data from multiple participants in an efficient way. The participants get the chance to locate their priorities, contrary opinions to other participants and, in addition, make use of communication approaches and other creative methods to reach other dimensions of the topic. (CDC 2008, Freitas et al. 1998, Remenyi 2011) So the interaction between the participants constitutes the most important difference by comparison to other types of interviews. (Kitzinger 1994)

Application in emergency management

This method was used in several research projects to collect information from a group of advisors or experts in a specific field. The group is sitting at a round table. The moderator asks open questions and leaves the discussion to the participants. In particular in European research projects different perspectives on processes, regulations or organisational structures with regard to the involved countries have to be discussed and considered. There is also the opportunity to use this interview method as a closing session after advisory board workshops to have a final participants-oriented discussion about the presented content.

Strengths and weaknesses of the method

This method is very helpful when there is a need for a larger number of interviews with different kind of experts. Furthermore, discussion helps to gain insights into the various perspectives of first responder or crisis manager to a specific topic. Focus groups are very popular, because the method is can be quickly and easily applied. But the most important advantage is the group dynamic emerging during the discussion. So information can be collected, which cannot be gathered through bilateral interactions with individuals. A challenge is that discussions can easily be dominated by a few individuals. Therefore, participant selection is very important. The analysis of the results is very time intensive and the results are always not representative. (CDC 2008, Freitas et al. 1998)

4.6.5 Guided expert interview

Interviews can be conducted in different ways. Interviews are a verbal conversation between the interviewer and interviewee. They can be used to find out about opinions, experiences, motivations and knowledge of the interviewee. Interviews can be conducted in a guided way (e.g. as scenario-based interviews), self-regulated or narrative interviews or focus groups. Focus groups use group dynamics to generate qualitative data.

Aim and applicability of the method

Experts are persons who have special knowledge and information about a specific field. They are frequently responsible for the development, implementation or control of solutions, strategies or policies at a company or agency. (Littig 2013, Meuser, Nagel 1991, Van Audenhove 2007) Expert interviews aim to get access to this specific information and expert knowledge relating to the underlying topic. Expert interviews are popular because they enable data collection about technical details or processes. (Littig 2013) here are three types of expert interviews that are aimed at different knowledge fields. Explorative expert interviews are conducted to obtain specific technical information, like details on operations, laws and influencing fields. The second expert interview type is the systematizing expert interview. This type focuses on process knowledge. The expert should be directly involved in the process. So that he provides information about routines, specific interactions and processes. It “aims at the reconstruction of expert’s special objective knowledge” (Van Audenhoven 2007). The third type is the theory-generating interview, which is conducted to obtain explanatory knowledge. The interviewer gathers information about the expert’s subjective interpretation of relevance, rules, beliefs and also ideas and ideologies and their inconsistencies. (Van Audenhoven 2007)

Before an expert interview starts the interviewer has to determine his role during the information exchange. He can decide to be a co-expert, expert outside the research field, a layperson, an authority (operator), a confederate with common normative background or a potential critic. (Littig 2013, Meuser, Nagel 1991, Van Audenhove 2007)

After the actual interview, data has to be processed and analysed. Six steps have to be conducted: transcription, paraphrasing, headlining, thematic comparison, scientific/ sociological conceptualisation and theoretical generalisation. (Bogner, Littig, Menz 2009, Littig 2013)

Application in emergency management

Scenario-based interviews are a specific form of expert interviews. This particular interview type requires sharing of expert knowledge regarding all three knowledge dimensions (technical, process and explanatory).

This interview type follows the specifications of a dedicated scenario and aligns questions in order to provide an interview flow based on realistic events. The aim of a scenario-based interview is to obtain specific information on the respective topic. (Remenyi 2011) The questions address all significant steps in a process. So the interviewer can obtain in-depth information around the general topic and gain further information about individual ways of dealing with problems in a given scenario.

Scenario-based interviews support understanding users’ application of a system or topic in their field. Furthermore, it is possible to reduce the complexity of the set-up during the interview.

Such interviews are often used in research projects. One example being the research project AirShield, which deals with the use of UAVs (unmanned air vehicles) to measure the spreading of dangerous substances in a three dimension way. In the evaluation phase it was the aim to gain valid responses to design questions and realisations from first responders and operation commands. This forces the research team to define a balance between real exercises but also to interview experts in a laboratory set-up without losing the context. In the context of AirShield this was done by scenario-based interviews including a dedicated software to simulate the flight of drones and spreading of substances in the overall AirShield-system.

This is a checklist to be used before starting an interview:

- Goals and research questions need to be defined from the project perspective.
- Overall scenario needs to be selected.
- Key factors need to be defined which mean a change in the reaction of first responder or operation command.
- Scenarios need to be elaborated including the flow / funnel of the scenario.
- Set-up for the interview and recordings need to be defined.

With a backup of all high-level topics it could be ensured to raise all relevant questions during the walkthrough of the scenario but also leaves space open to collect information about missing data, needs for other functions in that context.

Strengths and weaknesses of the method

Important for applying this method is a close contact to relevant stakeholder beforehand to build up realistic scenarios to raise the right questions during the interview. This method is a combination between the scenario analysis and a semi-structured interview mechanism. This forces the researcher to work out content for a realistic scenario and ensure to embed the research questions in the defined scenario. The method is useful especially at later stages of the project and is not used without prior knowledge about the working context of a specific emergency responder. In this process the experts are very important for the project, because they have valuable acquaintance of aggregated and specific knowledge (e.g. about processes, group behaviour, strategies). Furthermore, experts always network with other expert in their fields, so that the information exchanged and points of view of others can also be considered in the interview. By applying this method the interviewer gets information, which are difficult to obtain through other methods. But one has also to take into consideration that expert knowledge is not neutral. Therefore, it is a challenge to deal with anecdotal and illustrative information. (Van Audenhove 2007)

4.6.6 Criteria weighting

The application of MCA approaches for assessing systems' performances requires the weighting of the evaluation criteria selected. Various weighting methods are available, such as the Swing weight approach, point/budget allocation, ordinal ranking, pairwise comparison (from Saaty's Analytic Hierarchy Process), importance scales or importance-impact range graphs.

Point allocation seems to be the most intuitive approach. Users and/or decision makers are asked to allocate 100 points among the all criteria used for the evaluation.[24] This approach is based on the assumption that decision situations are quite similar to the allocation of a given (financial) budget, i.e. (financial) resources are spent on items in accordance with their relative importance. It is assumed that many users and/or decision makers are quite familiar with such kinds of allocation. In the literature it is conceded that this approach is easy to comprehend and apply. Still,

[24] For more information see OECD (2008, p. 96).

it is argued that weighting criteria without knowing their specific unit and score range can be inaccurate or - at worst - meaningless (Malczewski 1999).

Therefore, methods such as *Swing weight approach* are recommended, which also consider the range of the criteria values for determining the relative importance of criteria.[25] The rational of this this approach is that even though a criterion might be particularly relevant compared to another, if the performance scores of the alternatives assessed for this very relevant criterion do not substantially differ, these small criterion value differences should not have a disproportionately high impact on the final result. Applying the swing weight approach obviously does not exclude the opportunity that small differences in criterion scores intentionally have a big impact on the final result, if these are considered to be fundamentally important for the performance assessment.

So there is a trade-off between ease of application and accuracy of the weighting methods: Rating and ranking require less effort but lack the theoretical foundation of more sophisticated approaches such as the Swing weight method or pairwise comparisons, which in turn may lead to more precise results but are more laborious (Malczewski 1999). For practicability reasons we suggested to use the former approach to select suitable and feasible evaluation criteria and indicators and to apply the Swing weight approach for determining the relative weights for the aggregation of single performance aspects for the assessment of the overall performance of A4DEMOS compared to the existing disaster risk management systems.

4.7 Step 6: Data aggregation and analysis

4.7.1 PRIMATE

Data aggregation and analysis can be substantially facilitated by the use of software tools. The tool PRIMATE, which is an interactive software for **Pr**obab**i**listic **M**ulti-**At**tribute **E**valuation, allows for the comparative assessment of alternative options by means of CBA and/or MCA, considering different kinds of uncertainty.

The software's CBA module supports the identification of the most efficiently performing alternative on the basis of the options' net benefits[26]. Discounting has to be done prior to the use of PRIMATE.

The MCA module is based on the outranking concept PROMETHEE (**P**reference **R**anking **O**rganisation **Meth**od for **E**nrichment **E**valuations), which performs a pairwise comparison of all alternatives identified across all evaluation criteria. Alternatives, i.e. the baseline and the ex post scenario, form the rows and evaluation criteria form the columns of the evaluation matrix. In the first step the preferences of the users and/or decision makers, which were investigated in the data collection process to generate preference functions for each criterion, are used to transform the

[25] For a detailed description see e.g. Malczewski (1999), RPA (2004).

[26] More detailed information on CBA decision rules is provided in section 4.5.1.

differences in criteria values into preference values ranging from 0 to 1. On this basis a partial preference matrix is computed for each criterion. In the next step a total preference matrix (TPM) is computed as a weighted sum of the preference matrices of all criteria. Analysing the TPM the so-called leaving flows or outflow of an alternative can be computed by calculating the sum of the total preference values in each line of the TPM. This outflow can be interpreted as sum of "positive votes", i.e. the preferences for the respective alternative over the other option(s). Alternatively the sum of the so-called entering flows or inflow can be calculated as the sum of the column values of each alternative. This inflow can be interpreted as sum of "negative votes", i.e. the preferences of all other alternatives over the respective alternative.

When applying the decision rules of PROMETHEE I, alternatives are now ranked either based on the basis of their inflow and/or outflow. This means options are ordered from the one with the lowest/highest (=best) to the highest/lowest (=worst) inflow/outflow. If no incomparabilities between alternatives with regard to particular criteria exist rankings are perfectly invers.

If users and/or decision makers are interested in the net performance of an option, i.e. compensation of negative performance aspects by positive ones is allowed for, then the decision rule of PROMETHEE II can be applied. This version of the outranking concept ranks the options based on the net flows, i.e. the difference between outflow and inflow.

Figure 10 gives an overview of the application of the outranking concept PROMETHEE II.

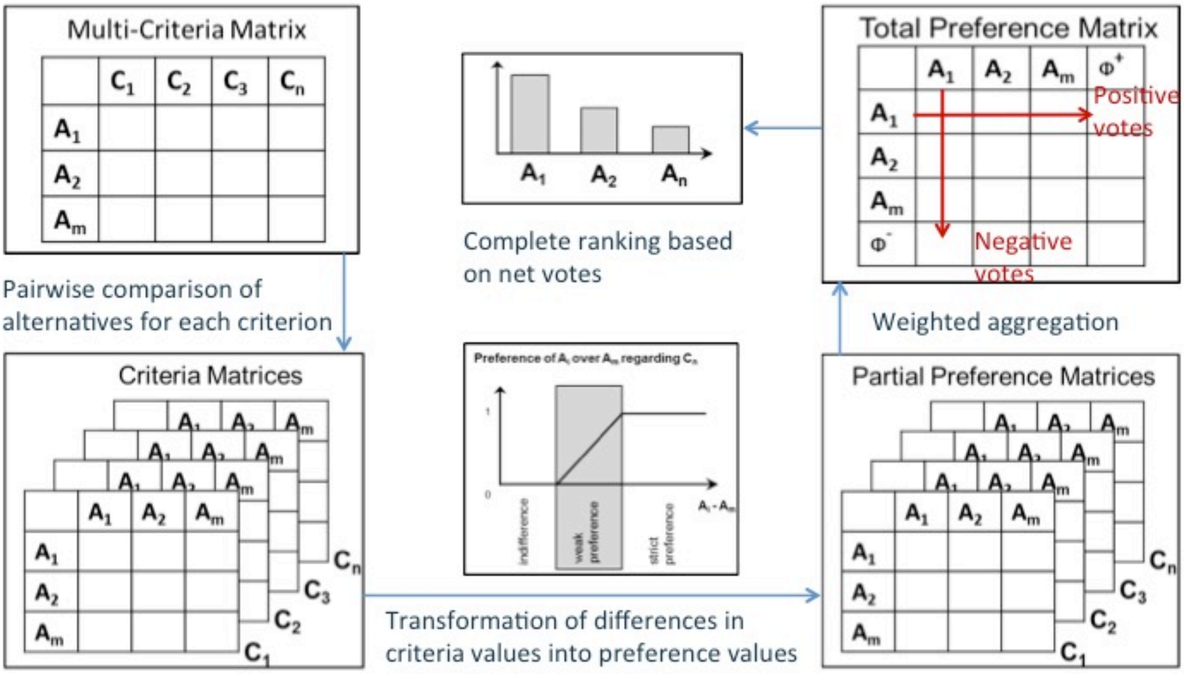

Figure 10: Overview of the application of the outranking concept PROMETHEE II
Note: Ai={A1,…,Am): Alternatives; Cj={C1,…,Cn}: Set of evaluation criteria; ϕ^-: Inflow; ϕ^+: Ouflow.

PRIMATE allows for the simultaneous and explicit consideration of the varying preferences of different users and/or decision makers for each criterion (= weights). Depending on the information at hand PRIMATE can use three different weighting methods.

- Global weighting, i.e. drawing of weights is completely random;
- Ordinal weighting, i.e. no assignment of specific weights, PRIMATE randomly samples numerical weights which are compatible with specified rank order specified by **one** user or decision maker;
- Cardinal weighting, i.e. assignment of precise numerical weights, e.g. by using the swing weight approach.

PRIMATE can consider uncertain information about alternatives‘ performances regarding single criteria probabilistically. The tool offers two ways to do that.

One possibility is to include uncertainty in the preference function. In this case PRIMATE uses a given range of criteria values to calculate a probability distribution for each alternative and each criterion. Then it performs pairwise comparison of these probability distributions. The rationale is that (1) if there is no overlap of these distributions then there is strict preference for the superior option and (2) if score ranges are overlapping but the mean value of one distribution is higher, depending on the specific preference function a preference value between 0 and 1 is assigned. This means that differences between options are transformed into preference values by the use of a stochastic preference function.

The second option is the use of Monte-Carlo simulations. This means that uncertainties are not included in the preference functions, but several PROMETHEE analyses are performed for a random sample of criterion values within a range to be defined.[27] This range can be specified as uniform distribution (minimum, maximum value), a triangular distribution (minimum, most likely, maximum value) or any other probability distribution, e.g. being the output of some model. PRIMATE randomly selects values out of the defined range and runs up to 10.000 MCAs. Results of all evaluations are then statistically analysed (arithmetic mean, standard deviation, ranking order) and documented. On this basis it can be stated that with a certain probability option A is the best performing alternative when considering all performance aspects and their relative importance as specified by the weighting set used.

If using the second option effects of the varying preferences and the uncertainty ranges of criteria values are documented in the final results. The results of the MCA are presented in PRIMATE in various ways. This includes not only the overall performance of an alternative considering all criteria and preferences but also its strengths or weaknesses with regard to a specific criterion or for specific users’ or decision makers’ preferences. This provides more insights in the evaluation results than the bare interpretation of ranking probabilities.

[27] For a guided tour offering comprehensive information on the methodical foundations of PRIMATE as well as practical advice for its use see Drechsler et al. (2009).

Summarizing the MCA with PRIMATE in a nutshell: The use of PRIMATE requires the identification of *alternative systems* to be compared and a set of *evaluation criteria* to be defined. The *preference functions* (indifference threshold, threshold of strict preference, shape of the preference function) for each user and/or decision maker and every criterion have to be specified[28]. The *weight* of every criterion has to be determined for each user and/or relevant decision maker in accordance with its relative importance. Data on the *performances* of the systems with regard to each evaluation criterion has to be collected. All this input data is entered into PRIMATE, which calculates rankings that provide probabilistic information about the best performing scenario, i.e. disaster risk management system.

Figure 11 gives an impression of the PRIMATE data matrix, treatment of uncertain input data, weighting functionalities as well as options to display evaluation results.

Figure 11: Main steps of the application of PRIMATE in the context of ANYWHERE

[28] For details see Drechsler et al. (2009, p. 3, p. 18-19).

5 References

Akter, T., and S. P. Simonovic, 2005: Aggregation of fuzzy views of a large number of stakeholders for multi-objective flood management decision-making. ***Journal of Environmental Management***, 77 (2), 133-143.

Alfieri, L., M. Berenguer, V. Knechtl, K. Liechti, D. Sempere-Torres, and M. Zappa, 2017: Flash Flood Forecasting Based on Rainfall Thresholds. In: Duan, Q., F. Pappenberger, J. Thielen, A. Wood, H.L. Cloke, and J.C. Schaake (eds): ***Handbook of Hydrometeorological Ensemble Forecasting***, Berlin, Heidelberg: Springer, 1-38.

Alfieri, L., F. Pappenberger, F. Wetterhall, T. Haiden, D. Richardson, and P. Salamon, 2014: Evaluation of ensemble streamflow predictions in Europe. ***Journal of Hydrology***, 517, 913-922.

Alfieri, L., D. Velasco, and J. Thielen, 2011: Flash flood detection through a multi-stage probabilistic warning system for heavy precipitation events. ***Advances in Geosciences***, 29, 69-75.

Arnal, L., H.L. Cloke, E. Stephens, F. Wetterhall, C. Prudhomme, J. Neumann, B. Krzeminski, and F. Pappenberger, 2017: Skilful seasonal forecasts of streamflow over Europe? ***Hydrology and Earth System Science Discussion Paper***.

Arnold, E., J. Clark, and M. Muscio, 2005: What the evaluation record tells us about Framework Programme Performance. ***Science and Public Policy***, 32, 385-397.

Bailey, K., 1994: ***Methods of social research***, New York: Free Press.

Bajpai, N., 2011: ***Business Research Methods***, Delhi: Pearson.

Bana E Costa, C. A., 1990: ***Reading in Multiple Criteria Decision Aid***, Berlin, Heidelberg, New York, Tokyo: Springer Verlag.

Bana E Costa, C. A., P.A. Da Silva, and F. Nunes Correia, 2004: Multicriteria Evaluation of Flood Control Measures: The Case of Ribeira do Livramento. ***Water Resources Management***, 18 (21), 263-283.

Bell, S., and S. Morse, 2003: ***Measuring Sustainability: Learning from doing***. London: Earthscan Publications.

Belton, V., and T.J. Stewart, 2002: ***Multiple Criteria Decision Analysis - An Integrated Approach***. Boston, Dordrecht, London: Kluwer.

Bengtsson, L., U. Andrae, T. Aspelien, Y. Batrak, J. Calvo, W. de Rooy, E. Gleeson, B. Hansen-Sass, M. Homleid, M. Hortal, K.-I. Ivarsson, G. Lenderink, S. Niemelä, K.P. Nielsen, J. Onvlee, L. Rontu, P. Samuelsson, D. Santos Muñoz, A. Subias, S. Tijm, V. Toll, X. Yang, and M. Ødegaard Køltzow, 2017: The HARMONIE–AROME Model Configuration in the ALADIN–HIRLAM NWP System. ***Monthly Weather Review***, 145 (5), 1919-1935.

Berenguer, M., and D. Sempere-Torres, 2013: Radar-based rainfall nowcasting at European scale: long-term evaluation and performance assessment. ***Preprints, 36th Conference on Radar Meteorology***, Breckenridge, USA: 15B.3.

Berenguer, M., C. Corral, R. Sanchez-Diezma, and D. Sempere-Torres, 2005: Hydrological validation of a radar-based nowcasting technique. ***Journal of Hydrometeorology***, 6, 532-549.

Berenguer, M., D. Sempere-Torres, and M. Hürlimann, 2015: Debris-flow forecasting at regional scale by combining susceptibility mapping and radar rainfall. ***Natural Hazards and Earth System Sciences***, 15, 587-602.

Berenguer, M., D. Sempere-Torres, and G.G.S. Pegram, 2011: SBMcast - An ensemble nowcasting technique to assess the uncertainty in rainfall forecasts by Lagrangian extrapolation, ***Journal of Hydrology***, 404, 226-240.

Bernard, R., 2000: ***Social Research Methods – Qualitative and Quantitative Approaches,*** Los Angeles, London, New Delhi, Singapore, Washington DC: SAGE Publications, 376-405.

Birch, S., and A. Gafni, 1992: Cost effectiveness/utility analyses. Do current decision rules lead us to where we want to be? ***Journal of Health Economics,*** 11 (3), 279-296.

Bogner, A., B. Littig, and W. Menz, 2009: ***Interviewing Experts***, Basingstoke: Palgrave Macmillan.

Brouwer, R., and J.M. Kind, 2005: Cost-benefit analysis and flood control policy in the Netherlands. Cost-benefit analysis and water resources management. In: Brouwer, R., and D. Pearce (eds.): ***Cost-Benefit Analysis and Water Resources Management***, Cheltenham: Edward Elgar, 71-91.

Brouwer, R., and R. van Ek, 2004: Integrated ecological, economic and social impact assessment of alternative flood control policies in the Netherlands. ***Ecological Economics***, 50 (1-2), 1-21.

Calheiros, R. V. and I. Zawadzki, 1987: Reflectivity-rain rate relationships for radar hydrology in Brazil. ***Journal of Climate and Applied Meteorology***, 26 (1), 118–132.

CDC, 2008: Data Collection for Program Evaluation: Focus Group. ***Evaluation Briefs***, No. 13. URL https://www.cdc.gov/healthyyouth/evaluation/pdf/brief13.pdf

Center for Innovation in Research and Teaching (CIRT), 2017: ***Types of Focus Groups***, URL https://cirt.gcu.edu/research/developmentresources/research_ready/focus_groups/types

Cloke, H.L., and F. Pappenberger, 2008: Evaluating forecasts of extreme events for hydrological applications: an approach for screening unfamiliar performance measures. ***Meteorological Applications***, 15, 181-197.

Corral, C., D. Velasco, D. Forcadell, and D. Sempere-Torres, 2009: Advances in radar-based flood warning systems. The EHIMI system and the experience in the Besòs flash-flood pilot basin. In: Samuels, P., S. Huntington, W. Allsop, and J. Harrop (eds): ***Flood Risk Management: Research and Practice***, Taylor & Francis, 1295-1303.

Corral, C., M. Berenguer, D. Sempere-Torres, L. Poletti , F. Silvestro, N. Rebora: ***Comparison of two systems for regional flash flood hazard assessment***, (in preparation).

Cover, J.A.G., 2005: Tabletop Role-Playing Games: Perspectives from Narrative, Game, and Rhetorical Theory. ***Journal of Chemical Information and Modeling,*** 53 (9), 1689–1699.

Crowther, D., and G. Lancaster, 2005: ***Research Methods – A concise introduction to research in management and business consultancy***, Oxford: Butterworth-Heinemann Ltd., 110-113.

Cucco, I., S. De Rosa, A. Passani, A. Georgiev, J. Pottebaum, C. Schäfer, T. Sauerland, K. Petersen, and I. Danilidis, I., 2016: ***SecInCoRe Deliverable 5.3, Validation strategy and first functional evaluation model of communication system concept.***

Di Giuseppe, F., C. Vitolo, B. Krzeminski: Fire danger: The forecast skill provided by ECMWF ensemble prediction system. ***Natural Hazards and Earth System Sciences*** (in preparation)

Di Giuseppe, F., F. Pappenberger, F. Wetterhall, B. Krzeminski, A. Camia, G. Libertá, and J. Miguel, 2016: The Potential Predictability of Fire Danger Provided by Numerical Weather Prediction. ***Journal of Applied Meteorology and Climatology***, 55 (11), 2469-2491.

Drechsler, M., 1999: Verfahren zur multikriteriellen Entscheidungsunterstützung bei Unsicherheit. In: Horsch H., and I. Ring (eds): ***Naturressourcenschutz und wirtschaftliche Entwicklung Nachhaltige Wasserbewirtschaftung und Landnutzung im Elbeeinzugsgebiet***, Leipzig: UFZ.

Drechsler, M., M. Lange, and V. Meyer, 2009: ***PRIMATE - An interactive software for Probabilistic Multi-Attribute Evaluation, Software handbook***, Leipzig: UFZ.

European Commission, 2009: ***Impact assessment guidelines. SEC 92.***

Fleming, I., 2017: URL http://www.sqa.net/iso9126.html.

Freitas, H., M. Oliveira, M. Jenkins, and O. Pojoy, 1998: ***The Focus Group, a qualitative research method***, Baltimore: Merrick School of Business, University of Baltimore.

Gowdy, J.M., 2009: ***Microeconomic Theory Old and New: A Student's Guide***, New Delhi: Orient BlackSwan.

Guy, G.B., and C.J. Kibert, 1998: Developing Indicators of Sustainability: US Experience. ***Building Research and Information***, 26 (1), 39-45.

Haiden, T., M. Janousek, J. Bidlot, L. Ferranti, F. Prates, F. Vitart, P. Bauer, and D.S. Richardson, 2016: Evaluation of ECMWF forecasts, including the 2016 resolution upgrade. ***ECMWF Technical Memorandum,*** 792.

Hanley, N., and C. Spash, 1996: ***Cost-Benefit Analysis and the Environment***, Cheltenham: Edward Elgar.

Hansjürgens, B., 2004: Economic valuation through cost-benefit analysis - possibilities and limitations. ***Toxicology***, 205 (3), 241-252.

Hewson, T., 2017: Use and verification of ECMWF products in Member and Co-operating States. ***ECMWF Technical Memorandum***, 797.

Hoppula, P., 2005: ***Tykkylumi ja otolliset säätilanteet sen aiheuttamille puustovaurioille***, Helsinki: Pro Gradu, Helsinki University.

Hürlimann, M., C. Abancó, J. Moya, M. Berenguer, and I. Vilajosana, 2016: Debris-Flow Monitoring for the Set-Up of a Warning and Alarm System -Experiences from the Pyrenees-. ***International Journal of Erosion Control Engineering***, 8, 107-113.

Hyogo Framework for Action (HFA) 2005-2015 (2007): ***Building the Resilience of Nations and Communities to Disasters Extract from the final report of the World Conference on Disaster Reduction*** (A/CONF.206/6).

Jolliffe, I.T., and D.B. Stephenson, 2012: ***Forecast Verification: A Practitioner's Guide in Atmospheric Science,*** Hoboken: John Wiley & Sons.

Keeney, R.L., and H. Raiffa, 1993: ***Decisions with Multiple Objectives - Preferences and Value Tradeoffs***, Cambridge: Cambridge University Press.

Kenyon, W., 2007: Evaluating flood risk management options in Scotland: A participant-led multi-criteria approach. ***Ecological Economics,*** 64, 70-81.

Kitzinger, J., 1994: The methodology of Focus Groups: the importance of interaction between research participants. ***Sociology of Health & Illness***, 16 (1), 103-121.

Klauer, B., M. Drechsler, and F. Messner, 2006: Multicriteria analysis under uncertainty with IANUS-method and empirical results. ***Environment and Planning C: Government and Policy***, 24, 235-256.

Lavaysse, C., J. Vogt, and F. Pappenberger, 2015: Early warning of drought in Europe using the monthly ensemble system from ECMWF, ***Hydrology and Earth System Sciences,*** 19, 3273-3286.

Littig, B., 2013: Expert Interviews. Methodology and Practice. In: ***IASR Lecture Series***, Vienna.

Malczewski, J., 1999: ***GIS and multicriteria decision analysis***, New York: John Wiley & Sons.

Meadows, D., 1998***: Indicators and information systems for sustainable development, Report to the Balaton Group***, Vermont: The Sustainability Institute.

Meißner, D., B. Klein, and M. Ionita, 2015: ***Towards a seasonal forecasting service for the German waterways—requirements, approaches, potential products***. Presentation at the HEPEX workshop on seasonal hydrological forecasting, 21-23rd September 2015, Norrköping, Sweden.

Messner, F., 2006: Applying participatory multicriteria methods to river basin management: Improving the implementation of the Water Framework Directive. ***Environment and Planning C: Government and Policy***, 24 (2), 159-167.

Meuser M., and U. Nagel, 1991: ExperInneninterviews – vielfach erprobt, wenig bedacht: ein Beitrag zur qualitativen Methodendiskussion. In: Garz D., and K. Kraimer (eds): ***Qualitativ-empirische Sozialforschung: Konzepte, Methoden, Analysen***. Opladen: Westdeutscher Verlag, 441-471.

Meyer V., R. Schwarze, N. Becker, V. Markantonis, J.C.J.M. van den Bergh, L.M. Bouwer, and C. Viavattene, 2014: Assessing the costs of natural hazards – state of the art and the way forward. In: Quevauviller, P. (eds): ***Hydrometeorological hazards: interfacing science and policy***. Chichester: John Wiley & Sons, 225-290.

Meyer, V., 2007: GIS-based Multicriteria Analysis as Decision Support in Flood Risk Management. ***UFZ Discussion Papers***, 6/2007, Leipzig: UFZ.

Meyer, V., S. Priest, S. and C. Kuhlicke, 2012: Economic evaluation of structural and non-structural flood risk management measures: examples from the Mulde River. ***Natural Hazards,*** 62, 301-324.

Ministry of Agriculture Fisheries and Food (MAFF), 1999: ***Flood and Coastal Defence Project Appraisal Guidance: Economic Appraisal***, London.

Monacciani, F., M. Navarra, A. Passani, and F. Bellini, 2011: ***D3.3a - SEQUOIA Final Self-Assessment Methodology.***

Munda, G., 1995: ***Multicriteria Evaluation in a Fuzzy Environment - Theory an Applications in Ecological Economics***, Heidelberg: Physica Verlag.

Nuottakari, J., K. Ylinen, J. Kauhanen, S. Neiglick, J. Ylläsjärvi, and J. Kilpinen, 2017: ***Multi-Model Ensemble Forecasts of Precipitation Type at European Airports***, 18th Conference on Aviation, Range, and Aerospace Meteorology, Seattle, Washington, USA.

OECD, 2008: ***Handbook on constructing composite indicators***, Paris: OECD.

Pappenberger, F., G. Jendritzky, H. Staiger, E. Dutra, F. Di Giuseppe, D.S. Richardson, and H.L. Cloke, 2015: Global forecasting of thermal health hazards: the skill of probabilistic predictions of the Universal Thermal Climate Index (UTCI). ***International Journal of Biometeorology,*** 59 (3), 311-323.

Pappenberger, F., M. Ramos, H.L. Cloke, F. Wetterhall, L. Alfieri, K. Bogner, A. Mueller, and P. Salamon, 2015: How do I know if my forecasts are better? Using benchmarks in hydrological ensemble prediction. ***Journal of Hydrology***, 522, 697-713.

Pappenberger, F., K. Scipal, and R. Buizza, 2008: Hydrological aspects of meteorological verification. ***Atmospheric Science Letters***, 9, 43-52.

Park, S., M. Berenguer, D. Sempere-Torres, C. Baugh, and P. Smith, 2017: Toward seamless high-resolution flash flood forecasting over Europe based on radar nowcasting and NWP: An evaluation with case studies. ***Geophysical Research Abstracts***, 19, 12158.

Park, S., M. Berenguer, D. Sempere-Torres, S. Pulkkinen, and J. Koistinen, 2017: Evaluation of the quality of precipitation nowcasting and hazard assessment in collaboration with Civil Protection agencies, ***ERICHA Deliverable E.1***, 26 pp.

Parker, D.J., 1991: ***The damage reducing effects of flood warnings.*** Report prepared for Halcrow national and for the national rivers authority (Anglian region) regional telemetry scheme appraisal, Middlesex: Flood Hazard Research Centre.

Parker, D.J., S.M. Tunstall, and S. McCarthy, 2007: New insights into the benefits of flood warnings: results from a household survey in England and Wales. ***Environmental Hazards,*** 7, 193-210.

Participation & Sustainable Development in Europe: Simulation Game, http://www.partizipation.at/simulationgame.html

Pearce D.J., and R. Smale, 2005: Appraising flood control investments in the UK. Cost-benefit analysis and water resorces management. In: Brouwer, R., and D.J. Pearce (eds): ***Cost-Benefit Analysis and Water Resources Management***. Cheltenham: Edward Elgar.

Penning-Rowsell, E.C., J.B. Chatterton, and D.J. Parker, 1978: **The Effect of Flood Warning on Flood Damage Reduction**. A report for the Central Water Planning Unit, Middlesex: Middlesex Polytechnic.

Pohjola, H., and A. Mäkelä, 2013: The comparison of GLD360 and EUCLID lightning location systems in Europe, ***Atmospheric Research,*** 123,117–128.

Prudhomme C., 2015: ***Operational seasonal hydrological forecasting in the UK***. Presentation at the HEPEX workshop on seasonal hydrological forecasting, 21–23rd September 2015, Norrköping, Sweden.

Pulkkinen, S., P. Rossi, M. Berenguer, and A. Mäkelä, 2015: Nowcasting of Lightning-related Hazards. **EDHIT Deliverable B.1,** 22 pp.

Remenyi, D., 2011: **Field Methods for Academic Research: Interviews, Focus Groups and Questionnaires**, Reading: Academic Conferences and Publishing International Limited, 49-52.

Rheinsberger, C., and H. Weck-Hannemann, 2007: Economic Approaches To Risk Evaluation, ***Working paper for the EU-MEDIN book project***, Zurich, 1-15.

Roche, C., 1999: ***Impact Assessment for Development Agencies: Learning to Value Change.*** Oxford, UK: Oxfam GB.

Rogers, E. M., 1983: ***Diffusion of Innovations (3rd ed.)***, New York: Free Press.

Rogers, E. M., 2003: ***Diffusion of innovations (5th ed.)***, New York: Free Press.

Rossi, P.J., V. Chandrasekar, V. Hasu, and D. Moisseev, 2015: Kalman filtering-based probabilistic nowcasting of object-oriented tracked convective storms. ***Journal of Atmospheric and Oceanic Technology***, 32 (3), 461-477.

Rossi, P.J., V. Hasu, J. Koistinen, D. Moisseev, A. Mäkelä, and E. Saltikoff, 2014: Analysis of a statistically initialized fuzzy logic scheme for classifying the severity of convective storms in Finland, ***Meteorological Applications***, 21 (3), 656-674.

RPA, 2004: ***Evaluating a multi-criteria analysis methodology for application to flood management and coastal defence appraisals***. R&D Technical Report. DEFRA.

Saltikoff, E., P. Lopez, A. Taskinen, and S. Pulkkinen, 2015: Comparison of quantitative snowfall estimates from weather radar, rain gauges and a numerical weather prediction model. ***Boreal Environment Research***, 20, 667-678.

Sekaran, U., and R. Bougie, 2013: ***Research Methods for Business: A Skill Building Approach***, Chichester: John Wiley & Sons.

Smith, P., F. Pappenberger, F. Wetterhall, J. Thielen, B. Krzeminski, P. Salamon, D. Muraro, M. Kalas, and C. Baugh, 2016: On the operational implementation of the European Flood Awareness System (EFAS). ***Flood Forecasting***, 313-348.

Sutanto, S.J., and H.A.J. Van Lanen, 2017: Exploration of drivers of the severe 1976, 1983, and 2003 droughts in Europe: An analysis with a climate model and an observation data set. ***Geophysical Research Abstracts***, 19, 7592.

Thöni M., A. Leiter, and H. Weck-Hannemann, 2009: Protective measures against natural hazards - Are they worth their costs?. In: Veulliet, E., J. Stötter J., and H. Weck-Hannemann (eds): ***Sustainable Natural Hazard Management in Alpine Environments***, 201-227.

Turner, R.K., D. Burgess, D. Hadley, E. Coombes, and N. Jackson, 2007: A Cost–benefit appraisal of coastal managed realignment policy. ***Global Environmental Change***, 17, 397-407.

Van Audenhove, L., 2007: ***Expert Interviews and Interview Techniques for Policy Analysis***, Brussel: Vrije University.

Van de Ven, A.H., 1991: The process of adopting innovations in organizations: Three cases of hospital innovations. In: ***People and technology in the workplace***, Washington: National Academy Press, 133-158.

Versini, P.A., M. Berenguer, C. Corral, and D. Sempere-Torres, 2014: An operational flood warning system for poorly gauged basins: demonstration in the Guadalhorce basin (Spain). ***Natural Hazards***, 71, 1355-1378.

Vincke, P., 1992: ***Multicriteria Decision-aid***, Chichester, New York, Brisbane, Toronto, Singapore: John Wiley & Sons.

Vitart, F., 2014: Evolution of ECMWF sub-seasonal forecast skill scores. ***Quarterly Journal of the Royal Meteorological Society Part B 114***, 1889-1899.

Vitolo, C., F. Di Giuseppe, and M. D'Andrea, 2017a: Caliver R package version 1.0. URL https://github.com/ecmwf/caliver, DOI: https://doi.org/10.5281/zenodo.376613.

Vitolo, C., F. Di Giuseppe, and M. D'Andrea, 2017b: Caliver: an R package for calibration and verification of forest fire gridded model outputs, PlosOne (under review).

Watkiss, P., A. Hunt, W. Blyth, and J. Dyszynski, 2014: The use of new economic decision support tools for adaptation assessment: A review of methods and applications, towards guidance on applicability. ***Climatic Change***, 132 (3), 401-416.

Young, R.A., 2005: Economic criteria for water allocation and valuation. Cost-benefit analysis and water resources management. In: Brouwer, R., and D.J. Pearce (eds): ***Cost-Benefit Analysis and Water Resources Management***. Cheltenham: Edward Elgar.

Zimmermann, H.J., and L. Gutsche, 1991: ***Multi-Criteria Analyse - Einführung in die Theorie der Entscheidungen bei Mehrfachzielsetzungen***, Berlin, Heidelberg, New York, Tokyo: Springer.

6 ANNEX

6.1 Validation measures to assess the different aspects of forecast quality

The Brier Score (BS) is the most common verification measure for assessing the accuracy of probability forecasts. The score is the mean squared error of the probability forecasts over the verification sample, BS is a so-called negatively oriented score, i.e. it ranges from 0 for a perfect forecast to 1 for the worst possible forecast outcome. Brier Scores of different verification samples should not be compared with each other unless the score is decomposed into three different terms (not shown here) representing the reliability, resolution and uncertainty attributes mentioned above, enabling a more detailed analysis of the verification results.

The Relative Operating Characteristic (ROC) is a widely used, popular method originating from signal detection theory to assess the performance of a (probabilistic) forecasting system to distinguish between the discrimination capability and the decision threshold of the system. The ROC curve is a graphical representation of the Hit rate (H; observed events that were correctly forecast) on the y-axis against the False Alarm Rate (F; number of false alarms given the event did not occur) along the x-axis for different potential decision thresholds. Graphically, the ROC curve is plotted from a set of probability forecasts by stepping a decision threshold (e.g. with 10% probability intervals) through the forecasts, and each probability decision threshold generating a 2*2 contingency table. Hence the probability forecast is transformed into a set of categorical "yes/no" forecasts and, consequently, a set of value pairs of H and F is obtained, forming the ROC curve. It is by default desirable that H is high and F is low. Therefore, the closer the point is to the upper left-hand corner of the graph, the better the forecast system. In a hypothetical perfect world there would be only correct forecasts with no false alarms, and a perfect forecast system would then be represented by a ROC "curve" that produces a single plot on the upper left-hand corner of the graph.

An attractive and widely used relative summary measure based on the ROC diagram is the ROC area (ROCA), which represents the area under the curve. ROCA is defined as =1 in a perfect forecast system and would decrease as the curve moves downward from the ideal top-left corner position. A useless forecast system with zero-skill is represented as a diagonal line, when H=F and, consequently, the ROC area is 0.5. Such a system cannot discriminate between occurrences and non-occurrences of the event.

The ROC (Relative Operating Characteristic) area (ROCA) and the Brier Score (BS) of the RAVAKE products (Section 2.1) were computed for the period 2007-2010. The accumulation period of precipitation was 12 hours. The forecasts are considered practically useful when the values of ROCA remain above 0.7. The results for the BS showed larger year-to-year variations than for the ROCA. The results indicated that the ECMWF EPS provides, in general, a solid basis for probabilistic precipitation forecasts and warnings. However, a calibration procedure is needed in future. These results indicate that the information from the ECMWF EPS system is useful on a daily level until 6-7 days ahead on average. The ROCA results showed that forecasts, with the given thresholds, are useful up to 3-4 days ahead, and these 6-hourly statistics indicate that the system is useful.

6.2 Additional tables

Table 7: Questionnaire for the co-evaluation of the co-production process.

WP/ROLE	
To which Work Package (WP) has your organisation contributed so far predominantly?	As an organisation we have so far contributed predominantly to the following WPs [please klick the WP you have been most involved in, multiple answer possible] WP 1 - Developing a framework for innovation co-ownership and co-evaluation ensuring the successful implementation of the project. WP 2 - Advanced forecasting models and tools to anticipate W&C event induced impacts WP 3 - Integrating impact models and tools in a Multi-Hazard operational early warning System WP 4 - Integrating impact models and tools in a Multi-Hazard operational early warning System WP 5 - Raising self-preparedness and self-protection to reduce population vulnerability WP 6 - User-driven pilot sites implementation and demonstration WP 7 - Innovation Exploitation, Business models and Market development WP 8 - Dissemination, communication, training and stakeholder engagement WP 9 - Project Management
To which WP do you personally contribute predominantly?	Personally, I have so far contributed predominantly to the following WPs [please klick the WPs you have been most involved in, multiple answer possible] WP 1 - Developing a framework for innovation co-ownership and co-evaluation ensuring the successful implementation of the project. WP 2 - Advanced forecasting models and tools to anticipate W&C event induced impacts WP 3 - Integrating impact models and tools in a Multi-Hazard operational early warning System WP 4 - Integrating impact models and tools in a Multi-Hazard operational early warning System WP 5 - Raising self-preparedness and self-protection to reduce population vulnerability WP 6 - User-driven pilot sites implementation and demonstration WP 7 - Innovation Exploitation, Business models and Market development WP 8 - Dissemination, communication, training and stakeholder engagement WP 9 - Project Management
Is your organisation leading a WP?	We are leading the following WP [please klick relevant WP (multiple answers possible] WP 1 - Developing a framework for innovation co-ownership and co-evaluation ensuring the successful implementation of the project. WP 2 - Advanced forecasting models and tools to anticipate W&C event induced impacts WP 3 - Integrating impact models and tools in a Multi-Hazard operational early warning System WP 4 - Integrating impact models and tools in a Multi-Hazard operational early warning System

	WP 5 - Raising self-preparedness and self-protection to reduce population vulnerability WP 6 - User-driven pilot sites implementation and demonstration WP 7 - Innovation Exploitation, Business models and Market development WP 8 - Dissemination, communication, training and stakeholder engagement WP 9 - Project Management
What is your role in the ANYWHERE-project?	I would describe myself primarily as a [klick relevant category; single answer]: - Researcher - Developer - User - Other, namely

Objectives of ANYWHERE	Strongly agree	Agree	Neither/nor	Disagree	Strongly disagree	Don't know
I am familiar with the overall objectives of the ANYWHERE project.						
I am familiar with the objectives of the WP(s) I am most involved in.						
In my opinion, we will achieve all objectives of the ANYWHERE project.						
In my opinion, we will achieve the objectives of the WP(s) I am mostly involved in.						
Please comment your answers, if you like.						

Personal role and contributions	Strongly agree	Agree	Neither/nor	Disagree	Strongly disagree	Don't know
I am aware of how I can contribute to the overall objectives of the ANYWHERE project.						
I am aware of how I can contribute to the objectives of the WP(s) I am most involved in.						
I have the feeling that I can contribute to make ANYWHERE a success.						
I would like to contribute more substantially to ANYWHERE.						

My motivation to contribute to the ANYWHERE project is high.						
Please comment your answers, if you like; for instance, which factors facilitate or complicate your engagement with the ANYWHERE project?	Comment field					

Information	Strongly agree	Agree	Neither/nor	Disagree	Strongly disagree	Don't know
I feel well informed about the progress of the ANYWHERE project.						
I feel well informed about the progress made in the WP(s) I am mostly involved in.						
To send project-related information I mainly use the following information channels	Comment field					
I think that there is sufficient room to reflect upon and discuss the progress of the ANYWHERE project.						
I would like to receive more information about … .	Comment field					
Can you please briefly evaluate the efficiency of the communication channels you are using to send and receive project-related information (e.g. Email/Newsletter; Revising documents on Basecamp; Virtual meetings)?	Comment field					
Which other communication channels would you like to use in future for which purposes and how often?	Comment field					
Please comment your answers, if you like.						

Pilot sites visits	Strongly agree	Agree	Neither/nor	Disagree	Strongly disagree	Don't know
Were you involved in any of the visits of the pilot sites?	Yes/No [if no, the survey is finished]; if yes, please indicate whether as pilot site partner / as visitor					
Please specify the pilot site your comments refer to:	Catalonia, Canton of Bern, Genoa, South Savo					
In my opinion the pilot site visits were very productive.						
The pilot sites visits met my expectations.						
For **pilot site partners**: I was able to express my opinion and my needs openly.						
For **pilot site partners**: I have the feeling, that others understood my opinion and my needs.						
For **pilot site partners**: I have the feeling, that the needs we expressed will be met by the ANYWHERE project.						
For **pilot site partners**: At this moment, I think it is very likely that I will operationally use ANYWHERE's products.						
For **pilot site visitors**: I have the feeling that I have a very good understanding of pilot site partners' needs.						
For **pilot site visitors**: I have the feeling that the project will be able to meet the needs expressed by pilot site partners.						
For both: We double-checked/ensured that the pilot site partners and developers have the same understanding of the pilot site needs.						
For **both**: Are you aware of any progress made since the pilot site visits.	Yes/no (if yes – In my opinion, very substantial progress/ some progress/ rather small progress.)					
For **both**: I can anticipate the advantages of using ANYWHERE's products in practice.						
Please comment your answers, if you like.						

Table 8: Overview impact areas, criteria and indicators of the ANYWHERE methodology

Impact area/effects	Criterion	Sub-criterion	Indicator
Economic impact			
Financial effects	False alarm-related cc		Total annual costs at your organisation due to false alarms triggered by the system.
	Procurement costs		Total onetime investment costs occurring at your organisation due to the implementation of the system.
	Operational costs		Total annual costs occurring at your organisation due to operation of the system.
	Maintenance costs		Total annual costs at your organisation for maintaining the system operational.
	Training costs (onetime)		Total onetime costs for training to enable end-users to operate the system.
	Training costs (annual)		Total annual costs for training to enable end-users to operate the system.
Economic effects	False alarm-related costs		Total annual costs due to false alarms triggered by the system (except costs occurring at your organisation).
	Damages occurred	Damages of building interior	Total annual/event-related damages of public/private buildings in the operational area.
		Damages of mobile assets	Total annual/event-related damages of public/private mobile assets in the operational area.
		Damage of buildings	Total annual/event-related damages of public/private buildings in the operational area.
		Damage of infrastructures	Total annual/event-related damages of public/private infrastructures in the operational area.
		Costs of business interruption	Total annual/event-related costs due to business interruption in the operational area.
	People affected	Fatalities	Total number of people who died due to the event.
		Casualties	Total number of people who were injured due to the event.

		Evacuated	Total number of people evacuated due to the event.
Technological impact			
Functionality*	Suitability		Functions and operations provide a reasonable and acceptable outcome.
	Accuracy		Frequency of incorrect or imprecise result caused by inadequate data.
			Frequency of incorrect or imprecise result caused by operation procedures.
	Interoperability		Number of functions, which are transferred easily between the system and other systems, or equipment, which is connected.
	Security		Frequency of occurrence of leak of secure output information or data.
			Frequency of occurrence of loss of important data.
			Frequency of occurrence of illegal access or illegal operation of the system.
Efficiency*	Time behaviour	Response time	Time to complete (a) specified task(s).
		Throughput	Number of tasks successfully performed over a given period of time
		Turnaround time	Waiting time between starting a (group of related) task(s) and their completion
		Waiting time	Proportion of time spent by users for the system to respond.
	Resource utilisation		Capacity of the system to perform required tasks using the given technical resources.
Maintainability*	Analysability		Effort to diagnose deficiencies or causes of failures of the system.
	Changeability		Effort to implement (a specified) modification(s) of the system.
	Stability		Frequency of occurrence of unexpected behaviour of the system after modifications.
	Testability		Effort to test the modified or non-modified system.
Usability*	Understandability		Ability of users to understand whether the system is suitable for their needs.
			Ability of the user to understand how a system can be used for (a) particular task(s).
	Learnability		Effort/time to learn how to use (particular functions of) the system.
			Effectiveness of help functionalities of the systems and of the documentation
	Operability	Suitability	Suitability of the system for the specified task(s).
		Self-descriptiveness	Self-descriptiveness of the system.

		Controllability	Controllability of the system.
		Conformity	Conformity of the system with user expectations.
		Error tolerance	Error tolerance of the system.
		Customizability	Possibility to customize the system.
	Attractiveness		Appearance of the system (design, colour etc.).
Reliability*	Maturity		Number of software failures.
	Fault tolerance		Capability of maintaining a specified performance level in cases of operation faults.
	Recoverability		Ability of the system to re-establish its adequate level of performance and recover the data directly affected in the case of a failure.
Portability*	Adaptability		Effort to adapt system to different specified environments.
	Installability		Effort to install the system in a user specific environment.
	Co-existence		Ability to use the system with other independent systems in a common environment sharing common resources.
	Replaceability		Effort to use the system in place of other specified system in the environment of that system.
Environmental impact			
Resource usage	Energy consumption		Energy used to operate the system per hour/day/annum
	Resource consumption	Paper	Quantity of paper used for operating the system per annum.
	Travel effort		Travel costs occurring for operating the system per annum.
System outputs	Waste	Technical equipment	Quantity of waste produced by operating the system per annum.
Social impact			
Effects on working routines and employment	Working routines		Time spent for compulsory activities under ordinary conditions/in case of an emergency.
			Level of satisfaction for execution of compulsory activities under ordinary conditions/in case of an emergency.
			Level of stress for execution of compulsory activities under ordinary conditions/in case of an emergency.
			Degree of rule-based decision-making.
			Degree of transparency of decision-making.

			Level of knowledge integration.
			Number of (optional) activities performed.
	Employment		Number of employees involved in the operation of the system.
Effects on knowledge production, use and sharing	Knowledge production	Inputs	Quantity of data input for decision making.
			Quality of data input for decision making.
		Processing	Ease of information processing under ordinary conditions/in case of an emergency.
		Outputs	Level of knowledge about relevant hazards.
			Level of knowledge about impacts of relevant hazards.
			Level of knowledge about relevant risks.
			Level of knowledge about management options.
	Knowledge use		Level of trust for using information provided by the system for decision-making under ordinary conditions/in case of an emergency.
	Knowledge sharing		Speed of hazard information dissemination to wider public/emergency forces in case of an emergency.
			Number of recipients of warnings.
			Ease of hazard information dissemination to wider public/emergency forces in case of an emergency.
			Ease of knowledge exchange with internal/external partners under ordinary conditions.
			Intensity of knowledge exchange with internal/external partners under ordinary conditions.
			Level of trust for exchanging information using the system.

Note: * Adaptation based on ISO/IEC 9126-2: Software Engineering – Product quality – Part 2: External metrics.

Table 9: Overview of criteria for assessment of software quality based on ISO/IEC 9126-2 and ISO/IEC 25010

ISO/IEC 9126-2		ISO/IEC 25010	
Functionality		Functional suitability	
	Suitability		Functional appropriateness
	Accuracy		Functional correctness
	Interoperability		Functional completeness
	Security		
Efficiency		Performance efficiency	
	Time behaviour		Time behaviour
	Resource utilisation		Resource utilisation
			Capacity
Maintainability		Maintainability	
	Analysability		Analysability
	Changeability		Modifiability
	Stability		Modularity
	Testability		Testability
			Reusability
Usability		Usability	
	Understandability		Appropriateness recognizability
	Learnability		Learnability
	Operability		Operability
	Attractiveness		User interface aesthetics
			User error protection
			Accessibility
Reliability		Reliability	
	Maturity		Maturity
	Fault tolerance		Fault tolerance
	Recoverability		Recoverability
			Availability
Portability		Portability	
	Adaptability		Adaptability
	Installability		Installability
	Co-existence		Replaceability
	Replaceability		
		Security	
			Confidentiality
			Integrity
			Non-repudiation
			Accountability
			Authenticity
		Compatibility	
			Co-existence
			Interoperability

Note: Red marked criteria were deleted; green marked criteria were added; blue marked criteria were renamed.