Project Number 700099

Call: H2020-DRS-01-2015

Project Title:

ANYWHERE

EnhANcing emergencY management and response to extreme WeatHER and climate Events

Subject:

Deliverable 3.3: Performance assessment of the MH-EWS platform on the Pilot Sites

Dissemination Level:

PU Public

Delivery date: **31st July 2018**

Month: **Month 26**

Organisation name of lead contractor for this deliverable: **RINA Consulting**

European Commission

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.

This document reflects only the authors' views and not those of the European Community. The information in this document is provided "as is" and no guarantee or warranty is given that the information is fit for any particular purpose. The user thereof uses the information at its sole risk and neither the European Community nor any member of the Consortium is liable for any use that may be made of the information.

Document Information

Title	**Performance assessment of the MH-EWS platform on the Pilot Sites**
Lead Author	Giovanni Napoli (RINA C.)
Contributors	Salvatore Ferraro (RINA C.) Ivan Tesfai (RINA C.) Gianni Loriga (RINA C.) Rafa Cuestas (HYDS) Alvaro Rodriguez (HYDS) Xavi Llort (HYDS) Claudia Di Napoli (UoR) Hannah Cloke (UoR) Fredrik Wetterhall (ECMWF) Claudia Vitolo (ECMWF) Estibaliz Gascon (ECMWF) Shaun Harrigan (ECMWF) Henny A.J. van Lanen (WUR) Samuel Jonson Sutanto (WUR) Remko Uijlenhoet (WUR) Ryan Teuling (WUR) Tomas Fernandez Montblanc (CFR) Paolo Ciavola (CFR) Marc Berenguer (UPC-CRAHI) Shinju Park (UPC-CRAHI) Alexandre Sanchez (UPC-CRAHI) Flavio Pignone (CIMA) von Lerber Annakaisa (FMI) Koistinen Jarmo (FMI) Bergman Tuomo (FMI)
Distribution	PUBLIC
Document Reference	D3.1 (Smith et al. 2017) D3.2 (Sancho et al. 2017)

Document History

Date	Revision	Prepared by	Organisation	Approved by
31.07.2018	V01	Giovanni Napoli	RINA C.	Gianni Loriga
26.09.2018	V02	Giovanni Napoli	RINA C.	Alexandre Sánchez

Related Documents

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

Summary

The main purpose of Work Package 3 (WP3) of the ANYWHERE project is to design and build the Multi-Hazard Early Warning System (MH-EWS), a flexible and scalable framework that integrates both forecast and impact models.

This document, “Performance assessment of the MH-EWS platform on the Pilot Sites”, presents the activities performed to validate the MH-EWS considering both the IT aspects of the platform and the tool/products there-in integrated.

This deliverable represents for the IT part the main functional assessment of the project, while for the performance evaluation of the model/product integrated in the MH-EWS it presents the assessment activities that have been just started and will be developed during the operational demonstration period at the Pilot Sites. The Deliverable D3.4 – “MH-EWS final operational prototype and final test results (M37)” will include the results.

Table of Contents

Definitions of the Acronyms

ACA	*Agència Catalana de l'Aigua* (Catalan Water Agency)
AEMET	*Agencia Estatal de Meteorologia*
API	Application Programming Interface
AUC	Area under the ROC Curve
BSS	Brier Skill Score
CFR	*Consorzio Futuro in Ricerca*
CIMA	*Centro Internazionale in Monitoraggio Ambientale*
CRPS	Continuous Rank Probability Score
CSI	Critical Success Index
ECMWF	European Centre for Medium-Range Weather Forecasts
EFAS	European Flood Awareness System
EFI	Extreme Forecast Index
ERIC	European Runoff Index
FAR	False Alarm Ratio
FF-EWS	Flash Flood Early Warning System (Flash flood hazard and impact assessment)
FMI	Finnish Meteorological Institute
FTP	File Transfer Protocol
FTPS	FTP Secure
FWI	Fire Weather Index
GeoTiff	Geostationary Earth Orbit Tagged Image File Format
HYDS	Hydrometeorological Innovative Solutions
IFS	Integrated Forecasting System
ISTIKE	East and South-East Finland Rescue Services Joint Situation and Coordination Centre
KPI	Key Performance Indicator
MAE	Mean Absolute Error
METAR	Meteorological Aerodrome Report
MH-EWS	Multi-Hazard Early Warning System
NBI	Normalized Bias
NetCDF	Network Common Data Form
NRMSE	Normalized Root Mean Square Error
NRT	Near Real-Time
NWP	Numerical Weather Prediction
OGC	Open Geospatial Consortium standard
OPERA	Operational Programme for the Exchange of Weather Radar Information
POD	Probability of Detection
PS	Pilot Site
QPE	Quantitative Precipitation Estimation
RAQ	Regional Air Quality
REST	Representational State Transfer
RINA C	RINA Consulting
RMSE	Root Mean Squared Error
ROC	Relative Operating Characteristic
SEDI	Symmetric Extremal Dependency Index

SIS2B	*Service Départemental d'Incendie et de Secours de Haute-Corse*
SOP	Standard Operating Procedures
SOS	Sensor Observation Service
SPI	Standardized Precipitation Index
SSH	Secure Shell
SSL	Secure Sockets Layer
SW	Software
SYNOP	Surface Synoptic Observations
TLS	Transport Layer Security
TWL	Total Water Level
UoR	University of Reading
UPC-CRAHI	*Universitat Politècnica de Catalunya – Centre de Recerca Aplicada en Hidrometeorologia* (Politechnic University of Catalonia – Center of Applied Research on Hydrometerology)
UTCI	Universal Thermal Climate Index
V&V	Validation & Verification
VPN	Virtual Private Network
VTM	Variable Threshold Method
WFS	Web Feature Service
WFS OGC	Web Feature Service
WHO	World Health Organization
WMO	World Meteorological Organization
WMS	Web Map Service
WSDL	Web Services Description Language
WUR	Wageningen University

1 Introduction

The main purpose of Work Package 3 (WP3) of the ANYWHERE project is to design and build the Multi-Hazard Early Warning System (MH-EWS), a flexible and scalable framework that integrates both forecast and impact models (developed within WP2).

A detailed description of the forecast and impact models and their requirements integrated (*connected* or *encapsulated*) into the MH-EWS have been compiled in Smith et al. (2017).

The second deliverable of the Work Package (Sancho et al. 2017) covers mainly the definition of the whole platform, the products generated by the integrated impact models, and the verification activities to be performed for the validation of the platform.

This document "Performance assessment of the MH-EWS platform on the Pilot Sites" presents the activities performed to validate the MH-EWS considering both the IT aspects of the platform and the tool/products there-in integrated.

The IT aspects of the platform, i.e. the functional and operational aspects, are evaluated in terms of platform accessibility, service provision via its interfaces and more in general the capability of the MH-EWS to provide the output of the models and tools that are encapsulated or linked according to the software and tests specifications of deliverables D3.1 and D3.2.

Once the MH-EWS platform is available, the performance evaluation of the models and tools integrated in the MH-EWS and developed in WP2 can start. According with the scope of work of the WP3 (and to avoid overlapping with other verification activities), the evaluation of the MH-EWS product/models within this work package is done on the algorithms/products that will be applied in the Pilot Sites. In fact, the European level is covered by WP2 and deliverable D2.5.

It is worth mentioning that this deliverable is issued a few months later than the Pilot Site implementation and several months before the end of the project; thus:

- It represents a verification point to assess the status of the MH-EWS platform respect to the proper implementation of the tool of WP4/WP5 and the Pilot Site (WP6);
- It will be followed (at M37 one year later) by the deliverable "D3.4 - MH-EWS final operational prototype and final test results (M37)" that will include the results of the evaluation methodology defined in this deliverable D3.3 along the duration of the demonstrations in the Pilot Sites.

Therefore, considering what above, this deliverable represents for the IT part the main functional assessment of the project, while for the performance evaluation of the model/product integrated in the MH-EWS it presents the assessment activities that will be developed during the Pilot Site operational demonstration period. The Deliverable D3.4 – "MH-EWS final operational prototype and final test results (M37)" will include the results.

1.1 Objective

This document covers the evaluation of the MH-EWS focusing both on:

- the IT aspects according to the software and test specification of D3.1 and D3.2; and
- the performance of the models and tools (developed in WP2) applied in real time in the Pilot Sites.

The second point of the list above means that this deliverable presents the methodologies for the assessment of the model and tools developed within the WP2 and integrated in the Pilot Site via the MH-EWS, namely:

- Meteorological Forecasts and Nowcasts products (ECMWF, CIMA, FMI);
- Floods, Flash-Floods and landslides Products (ECMWF, UPC-CRAHI, CIMA);
- Storm Surges products (CFR);
- Air Quality and health products (UoR);
- Fire products (ECMWF, CIMA);
- Droughts products (WUR).

The final results of the verification activities will be included "D3.4 - MH-EWS final operational prototype and final test results (M37)" that will include the results of the evaluation methodology defined in this deliverable along the duration of the demonstrations in the Pilot Sites.

1.2 Organization of the document

The following of the document is organized in a few main chapters:

- Chapter 2 introduces to the strategy of the validation defined within the WP3 as described to D3.2 (this chapter updates the presentation of the strategy proposed in D3.2 according to the project development) and, moreover, it presents the interfaces of the verification activities presented in this deliverable respect to other verifications in the project;
- Chapter 3 presents the SW Validation in terms of main modules and interfaces including the: tests on the single modules of the MH-HWS; use cases tests that address the MH-HWS performances;
- Chapter 4 presents the validation activities for the algorithms and related products through Pilot Site experience.

2 Validation and Verification Strategy

The services, interfaces and products verified and validated are those defined in Deliverable D3.1 and D3.2.

The V&V activities correspond to Task 3.4 of the project, and, as described in details in D3.2, they are structured in three phases that run in parallel with the MH-EWS development. The first two phases are focused on the MH-EWS platform testing (mainly functional testing on the IT part), the third aims at the models/tools refining (performance testing of the platform). The strategy applied, although it is in line with that presented in D3.2, it needed a tuning during the project development:

- First Phase: Validation of the MH-EWS framework integrating both the forecast and impact models (WP2), and the existing Pan-European platforms as standalone, i.e. proper availability of the modules of the platform before the Pilot Site implementation;
- Second Phase: Validation of the services and interfaces functionality as used by the solution developed for the self-preparedness and self-protection tools (together indicated as self-p*). This part is covered by use case tests;
- Third Phase: Validation of the impact prediction capabilities and tools developed by the WP2 partners in the Pilot Sites (WP6).

2.1 First and Second Phase Testing

First and Second Phase Testing activities evaluate the IT Part of the MH-EWS from the functional point of view. While the First Phase is strictly linked to the development phase and functional test on the single modules, the second is linked to the integration phase and uses of the whole running platform:

- The first phase of verification is carried out mainly by the software developers and it starts with the software development to verify the integration of existing and improved impact forecasting algorithms in the platform (once it has been encapsulated or externally linked) and single modules.
- The second phase is carried out to verify that the MH-EWS provides the output to its users as expected, i.e. all the components interact between them properly to provide some specific results that will be used by the Pilot Site.

The partners dealing with the Pilot Site are both those of WP2 that provide products and models for impacts assessment and WP4 and WP5 tool developers that use the MH-EWS output for their services.

Namely, the outputs from WP4 partners (users of the MH-EWS) are:

- Different versions of the A4* platforms;
- Risk Assessment tool;
- Local transport and logistics management tool;

- Common Picture via communication module.

The interaction between the MH-EWS and the self-p* tools of WP5 will be verified by the tools developers as for their use:

- Tool for reducing storm-driven impacts on electricity transmission grids;
- Tool for enabling self-awareness and self-response of the logistic platforms of the food distribution companies during snowfall episodes;
- Tool for increasing self-protection in camping sites located in flood prone areas;
- Increasing self-awareness and self-protection in front of flooding risk in schools.

Chapter 3 presents the activities done for the first two phases. An official verification of the functionality of the interfaces of the system and its capability to provide data from external sources and models has been performed to achieve the Milestone M3.2: "MH-EWS first operational prototype ready for implementation on the Pilot Sites (M19)".

2.2 Third Phase Testing

The Third Phase Testing focuses on the evaluation of the performance of the impact forecasting products provided by the MH-EWS, in terms of the products developed over the WP2 algorithms in relation to the Pilot Site needs. This testing phase started with the MH-EWS implementation on the Pilot Sites and therefore this document presents the methodology and preliminary results.

Chapter 7 presents this assessment that have been developed directly by the responsible of the products based on the following general rules that have been defined within the T3.4:

- Evaluation of the impact forecasting accuracy through the comparison of the estimation made by the models with the occurrence, and/or evaluation with respect to the emergency actions triggered.
- Evaluation of impact forecasting using reference data, if available.
- Comparison of the reliability/usability and importance/weight of using the ANYWHERE MH-EWS in front of other tools already used in site.
- Satisfaction of the final users according to their needs in terms of assets included in the impact estimation. The users could be interviewed to understand if the products of MH-EWS fulfil their needs of emergency management and self-p*.
- Evaluation of the response time. Capability of the MH-EWS to produce the outputs with enough time in advance.

2.3 Interfaces Toward Other Project Activities

Within the ANYWHERE project there are several verification activities to properly drive the project respect to the actual capability and specifications.

To integrate the effort of all the involved partners and avoid any overlapping among the diverse WPs, the WP leaders performed a coordination process that brought to:

- In WP1 there are KPI for the verification of the ANYWHERE project as a whole. Respect to these activities, the WP3 activities do not present overlaps because they are technical and linked to the platform and single product/model. Of course, the proper implementation of WP3 brings to benefit to the whole project success;
- In WP2 there are evaluation tasks related to the algorithms. Specifically, the deliverable D2.5 "Uncertainty and robustness assessment" (M37) will focus on the evaluation of the performance of the algorithms of the MH-EWS at European scale (while the evaluation of WP3 would be done at the PSs);
- In WP3 the validation activities are related to the MH-EWS as platform and it's products/models applied in the Pilot Sites.

It's worth nothing that some of WP1's KPIs, in particular those known as "System-Related Indicators", will be considered as a reference for the definition of the KPIs related to the MH-EWS. MH-EWS KPIs are considered in performance tests definition during the project development, especially in the Third Phase Testing, product/model performance evaluation.

The validation of the models and tool provided by WP2 will continue during the Pilot Sites' implementations (in WP6, until M38). Thus, the tests results will provide feedback to the models/tools developers in order to refine and improve them.

3 MH-EWS Platform Evaluation (IT Part)

Along with the software development, all interfaces of the MH-EWS have been evaluated internally by the technical team. A testing section that involved not only the software developers, but also the verification manager team was carried out for the Milestone M3.2 at the end of December 2017. That activity has brought to the following results, as described in the following paragraphs.

It is worth noting that the validation activities of the IT platform will be intrinsically continued during the performance evaluation of the tools and products, considering that they use the IT infrastructure for assessment. Thus, the evaluation has been of critical importance for the starting of the Pilot Sites and tools assessment.

3.1 Introduction

The aim of this section is to validate that the MH-EWS provides the expected results and they are provided under some performance constraints. Thus, it is important not only that the results are as expected but also that they are provided on time and with controlled amount of resources.

The MH-EWS assessment has been therefore divided in two parts:

- First, functional tests of the MH-EWS have been done. These functional tests focus on the Gateway and Data supply modules, which interact with other systems. The following Figure 1 presents the general architecture of the MH-EWS.

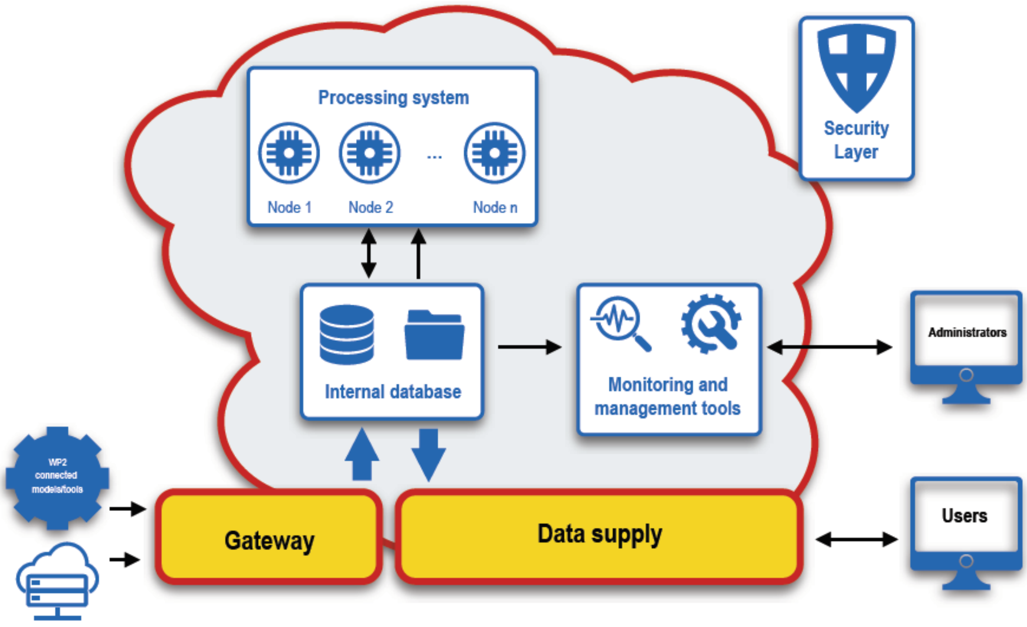


Figure 1: General architecture of the MH-EWS. The modules tested in the functional tests are marked in yellow.

The objective of the functional tests is to check that the modules provide the expected results in specific tests. Thus, these functional tests can be considered as a continuation of the validation presented in the Milestone 3.2 (MS7).

- Second, a performance assessment of several use cases, where the MH-EWS components interact between them to provide some specific results. Here, the goal is not only to check whether the results are the expected ones but also that they are calculated or generated under certain indicators (time spent, resources consumed, etc.). These use cases are based on the requirements and specifications defined in Deliverable 3.2.

The following sections present both functional tests and performance assessment carried out with the MH-EWS. All of them have been run in the server where the MH-EWS is deployed, which has the following specifications:

- CPU: 48 cores @ 2.3 GHz
- 64Gb RAM
- 28Tb HDD
- This server can be accessed remotely via SSH protocol and VPN.

3.2 Functional tests

This section presents the different functional tests carried out to test the MH-EWS. The aim of these tests was to confirm the different functionalities described in Deliverable 3.2 are working and that provide the expected results.

A first set of functional tests was presented and performed in Milestone 3.2 to demonstrate that the MH-EWS was ready for the implementation on the Pilot Sites. Therefore, the functional tests presented in this document are an update to test that all the functionalities are still working as expected. These tests are focused on those MH-EWS services and interfaces offering data to the users (detailed information of all the services and interfaces is presented in Deliverable 3.2.), the rest of services and interfaces are validated with the use case tests described in section 3.3.

The following sub-sections present the tests carried out on the different MH-EWS services and interfaces.

3.2.1 Products' Catalogue service

The Products' Catalogue service provides meta-information (not the data) about the products offered by the MH-EWS and it is a service intended for machine-to-machine interaction and it provides a secure REST API. The design of the products' catalogue service is included in the Deliverable 3.2 (Section 2.5.4.2, page 43).

3.2.1.1 Tests definition

This section presents the different tests carried out to validate the Products' Catalogue service, describing each of them, how to pass them and the expected result.

In general, the validation activities for the REST APIs will focus on the following:

- Correct use of the API key mechanism (where foreseen) used for the users' authentication and authorization (Security Layer) for accessing resources (through the GET, PUT, DELETE and POST HTTP commands used in REST).
- Extensive testing of the complete REST web services, considering all the possible output formats (e.g. JSON, XML, etc.).

The specific validation activities for the Products' Catalogue service consist of the testing of the correct access to the API of the service to retrieve data from the server (GET method), and the testing of the response format (both in the case of JSON and XML). The following sections present the different tests to be passed. All of them can be passed, for example, using HURL[1], an online tool to make HTTP requests or using ad hoc scripts (using Python or other languages).

The different tests were passed by several partners during the achievement of the Milestone 3.2 in December 2017. Each partner had a specific API-key, to confirm that the service provides different information for each user.

Get catalogue information

This test validates the service can provide products information for a specific user.

Table 1: Get catalogue information test.

URL	https://rest.mhews.anywhere-h2020.eu/v1/prod_catalogue
Method	GET
Headers	X-API-KEY: <YOUR_API_KEY>
Expected result	[{ "name":"ifs_hres_10m_u_wind_speed", "description":"10 meters U wind speed", "unit":"m/s", "id_product":"10m_u_wind_speed", "time_step":"10800", "update_frequency":"604800", "first_data":"2017-12-05 18:00:00", "last_data":"2017-12-24 00:00:00", "data_type":"raster", "bounding_box":"[-27,33,45,73.5]", "resolution":"[0.1,0.1]", "available_formats":"raster" }, { "name":"ifs_hres_10m_v_wind_speed", "description":"10 meters V wind speed", "unit":"m/s", "id_product":"10m_v_wind_speed", "time_step":"10800", "update_frequency":"604800", "first_data":"2017-12-05 18:00:00", "last_data":"2017-12-24 00:00:00", "data_type":"raster", "bounding_box":"[-27,33,45,73.5]", "resolution":"[0.1,0.1]", "available_formats":"raster" }, { "name":"ifs_hres_2m_temperature",

[1] HURL webpage: https://www.hurl.it/

```
        "description":"2 meters temperature",
        "unit":"°K",
        "id_product":"2m_temperature",
        "time_step":"10800",
        "update_frequency":"604800",
        "first_data":"2017-12-05 18:00:00",
        "last_data":"2017-12-24 00:00:00",
        "data_type":"raster",
        "bounding_box":"[-27,33,45,73.5]",
        "resolution":"[0.1,0.1]",
        "available_formats":"raster"
    },
    {
        "name":"ifs_hres_dew_point_temperature",
        "description":"Dew point temperature",
        "unit":"°K",
        "id_product":"dew_point_temperature",
        "time_step":"10800",
        "update_frequency":"604800",
        "first_data":"2017-12-05 18:00:00",
        "last_data":"2017-12-24 00:00:00",
        "data_type":"raster",
        "bounding_box":"[-27,33,45,73.5]",
        "resolution":"[0.1,0.1]",
        "available_formats":"raster"
    },
    {
        "name":"ifs_hres_precipitation",
        "description":"Precipitation",
        "unit":"mm",
        "id_product":"precipitation",
        "time_step":"86400",
        "update_frequency":"604800",
        "first_data":"2017-12-05 18:00:00",
        "last_data":"2017-12-24 00:00:00",
        "data_type":"raster",
        "bounding_box":"[-27,33,45,73.5]",
        "resolution":"[0.1,0.1]",
        "available_formats":"raster"
    }
]
```

Wrong HTTP method

This test validates that the service does not return a proper response when making requests to no-GET method.

Table 2: Wrong HTTP method test.

URL	https://rest.mhews.anywhere-h2020.eu/v1/prod_catalogue
Method	POST
Headers	X-API-KEY: <YOUR_API_KEY>
Expected result	HTTP error 500: Internal server error

No API-KEY

This test validates that the service returns an error when no API key is specified.

Table 3: No API-KEY test.

URL	https://rest.mhews.anywhere-h2020.eu/v1/prod_catalogue
Method	GET
Expected result	HTTP error 401: Not defined api-key

Wrong API-KEY

This test validates that the service return an error when the specified API key does not belong to a defined user.

Table 4: Wrong API-KEY test.

URL	https://rest.mhews.anywhere-h2020.eu/v1/prod_catalogue
Method	GET
Headers	X-API-KEY: 1234567890
Expected result	HTTP error 401: Error in authentication

XML response

This test validates that the service can provide products information in XML format.

Table 5: XML response test.

<table>
<tr><td>URL</td><td>https://rest.mhews.anywhere-h2020.eu/v1/prod_catalogue</td></tr>
<tr><td>Method</td><td>GET</td></tr>
<tr><td>Headers</td><td>X-API-KEY: <YOUR_API_KEY>
Accept: application/xml</td></tr>
<tr><td>Expected result</td><td><pre>
<?xml version="1.0" encoding="UTF-8"?>
<products>
    <product>
        <name>ifs_hres_10m_u_wind_speed</name>
        <description>10 meters U wind speed</description>
        <unit>m/s</unit>
        <idProduct>10m_u_wind_speed</idProduct>
        <timeStep>10800</timeStep>
        <updateFrequency>604800</updateFrequency>
        <firstData>2017-12-05 18:00:00</firstData>
        <lastData>2017-12-24 00:00:00</lastData>
        <dataType>raster</dataType>
        <availableFormats>raster</availableFormats>
        <boundingBox>
            <left>-27</left>
            <bottom>33</bottom>
            <right>45</right>
            <top>73.5</top>
        </boundingBox>
        <resolution>
            <x>0.1</x>
            <y>0.1</y>
        </resolution>
    </product>
    <product>
        <name>ifs_hres_10m_v_wind_speed</name>
        <description>10 meters V wind speed</description>
        <unit>m/s</unit>
        <idProduct>10m_v_wind_speed</idProduct>
        <timeStep>10800</timeStep>
        <updateFrequency>604800</updateFrequency>
        <firstData>2017-12-05 18:00:00</firstData>
        <lastData>2017-12-24 00:00:00</lastData>
        <dataType>raster</dataType>
        <availableFormats>raster</availableFormats>
        <boundingBox>
            <left>-27</left>
            <bottom>33</bottom>
            <right>45</right>
            <top>73.5</top>
        </boundingBox>
        <resolution>
            <x>0.1</x>
            <y>0.1</y>
</pre></td></tr>
</table>

```
        </resolution>
    </product>
    <product>
        <name>ifs_hres_2m_temperature</name>
        <description>2 meters temperature</description>
        <unit>°K</unit>
        <idProduct>2m_temperature</idProduct>
        <timeStep>10800</timeStep>
        <updateFrequency>604800</updateFrequency>
        <firstData>2017-12-05 18:00:00</firstData>
        <lastData>2017-12-24 00:00:00</lastData>
        <dataType>raster</dataType>
        <availableFormats>raster</availableFormats>
        <boundingBox>
            <left>-27</left>
            <bottom>33</bottom>
            <right>45</right>
            <top>73.5</top>
        </boundingBox>
        <resolution>
            <x>0.1</x>
            <y>0.1</y>
        </resolution>
    </product>
    <product>
        <name>ifs_hres_dew_point_temperature</name>
        <description>Dew point temperature</description>
        <unit>°K</unit>
        <idProduct>dew_point_temperature</idProduct>
        <timeStep>10800</timeStep>
        <updateFrequency>604800</updateFrequency>
        <firstData>2017-12-05 18:00:00</firstData>
        <lastData>2017-12-24 00:00:00</lastData>
        <dataType>raster</dataType>
        <availableFormats>raster</availableFormats>
        <boundingBox>
            <left>-27</left>
            <bottom>33</bottom>
            <right>45</right>
            <top>73.5</top>
        </boundingBox>
        <resolution>
            <x>0.1</x>
            <y>0.1</y>
        </resolution>
    </product>
    <product>
        <name>ifs_hres_precipitation</name>
        <description>Precipitation</description>
        <unit>mm</unit>
        <idProduct>precipitation</idProduct>
        <timeStep>86400</timeStep>
        <updateFrequency>604800</updateFrequency>
        <firstData>2017-12-05 18:00:00</firstData>
        <lastData>2017-12-24 00:00:00</lastData>
        <dataType>raster</dataType>
        <availableFormats>raster</availableFormats>
        <boundingBox>
            <left>-27</left>
            <bottom>33</bottom>
            <right>45</right>
            <top>73.5</top>
        </boundingBox>
        <resolution>
            <x>0.1</x>
            <y>0.1</y>
        </resolution>
    </product>
</products>
```

3.2.1.2 Tests results

The following table depicts the results obtained for the different tests:

Table 6: Summary table of the MH-EWS tests related to the Products' Catalogue service.

Test	CRAHI	CIMA	AIRBUS	RINAC	HYDS
Get Catalogue information	OK	OK	OK	OK	OK
Wrong HTTP method	OK	OK	OK	OK	OK
No API-KEY	OK	OK	OK	OK	OK
Wrong API-KEY	OK	OK	OK	OK	OK
XML response	OK	OK	OK	OK	OK

3.2.1.3 Evaluation

All the functionalities are working as expected, according to the results. Thus, no additional improvements are considered at this point.

3.2.2 Bulk data service

The Bulk data service allows the user to retrieve bulk data from the MH-EWS (that is, raster data fields or geospatial vector data), served as regular files.

3.2.2.1 Tests definition

The tests carried out to validate the bulk data service consist in checking the correct access to the API of the service to retrieve the requested data. The following sections presents the different tests performed. All these tests contain an example of how to run them using the CURL command in bash.

These tests were initially passed by several partners during the achievement of the Milestone 3.2 in December 2017. Each partner had a specific API-KEY and product to make the requests. The same tests were passed by the same partners to confirm whether the results were the same or not.

Get files

This test validates the bulk data service returns the requested files.

Table 7: Get files test.

URL	https://rest.mhews.anywhere-h2020.eu/v1/bulkd_data?id_product=<PRODUCT_NAME>&date=<DATE>
Method	GET
Headers	X-API-KEY: <YOUR_API_KEY>
CURL instruction	`curl -X GET -H "X-API-KEY: <YOUR_API_KEY>" "https://rest.mhews.anywhere-h2020.eu/v1/bulk_data?id_product=<PRODUCT_NAME>&date=<DATE>" -o data.zip`
Expected result	A zip file including the NetCDF file for the given date.

Wrong HTTP method

This test validates that the bulk data service triggers and error when the used HTTP method is not GET.

Table 8: Wrong HTTP method test.

URL	https://rest.mhews.anywhere-h2020.eu/v1/bulkd_data?id_product=<PRODUCT_NAME>&date=<DATE>
Method	DELETE
Headers	X-API-KEY: <YOUR_API_KEY>
CURL instruction	`curl -X DELETE -v -H "X-API-KEY: <YOUR_API_KEY>" "https://rest.mhews.anywhere-h2020.eu/v1/bulk_data?id_product=<PRODUCT_NAME>&date=<DATE>" -o data.zip`
Expected result	HTTP error 500: Interval server error

No API-KEY

This test validates that the service returns an error when no API-KEY is specified.

Table 9: No API-KEY test.

URL	https://rest.mhews.anywhere-h2020.eu/v1/bulkd_data?id_product=<PRODUCT_NAME>&date=<DATE>
Headers	GET
CURL instruction	`curl -X GET -v "https://rest.mhews.anywhere-h2020.eu/v1/bulk_data?id_product=<PRODUCT_NAME>&date=<DATE>" -o data.zip`
Expected result	HTTP error 401: Not defined api-key

Wrong API-KEY

This test validates that the bulk data service triggers an error in case API-KEY does not belong to a known user.

Table 10: Wrong API-KEY test.

URL	https://rest.mhews.anywhere-h2020.eu/v1/bulkd_data?id_product=<PRODUCT_NAME>&date=<DATE>
Method	GET
Headers	X-API-KEY: 1234567890
CURL instruction	`curl -X GET -H "X-API-KEY: 1234567890" "https://rest.mhews.anywhere-h2020.eu/v1/bulk_data?id_product=<PRODUCT_NAME>&date=<DATE>" -o data.zip`
Expected result	HTTP error 401

Get files of a non-existing product

This test validates that the service returns an error when requesting a product not defined in the MH-EWS.

Table 11: get files of a non-existing product test.

URL	https://rest.mhews.anywhere-h2020.eu/v1/bulkd_data?id_product=<PRODUCT_NAME>&date=<DATE>
Method	GET
Headers	X-API-KEY: <YOUR_API_KEY>
CURL instruction	`curl -X GET -H "X-API-KEY: <YOUR_API_KEY>" "https://rest.mhews.anywhere-h2020.eu/v1/bulk_data?id_product=<PRODUCT_NAME>&date=<DATE>" -o data.zip`
Expected result	HTTP error 401

Get files of a non-contracted product

This test validates that the service returns an error when requesting a non-contracted product.

Table 12: Get files of a non-contracted product test.

URL	https://rest.mhews.anywhere-h2020.eu/v1/bulkd_data?id_product=<PRODUCT_NAME>&date=<DATE>
Method	GET
Headers	X-API-KEY: <YOUR_API_KEY>
CURL instruction	`curl -X GET -H "X-API-KEY: <YOUR_API_KEY>" "https://rest.mhews.anywhere-h2020.eu/v1/bulk_data?id_product=<PRODUCT_NAME>&date=<DATE>" -o data.zip`
Expected result	HTTP error 401

Get files without a date

This test validates that the service returns an error when request data without specifying the date (mandatory).

Table 13: Get files without a date test.

URL	https://rest.mhews.anywhere-h2020.eu/v1/bulkd_data?id_product=<PRODUCT_NAME
Method	GET
Headers	X-API-KEY: <YOUR_API_KEY>
CURL instruction	`curl -X GET -H "X-API-KEY: <YOUR_API_KEY>" "https://rest.mhews.anywhere-h2020.eu/v1/bulk_data?id_product=<PRODUCT_NAME>" -o data.zip`
Expected result	HTTP error 500

Get files of a time interval

This test validates that the service returns data of a defined time interval.

Table 14: Get files of a time interval test.

Endpoint	https://rest.mhews.anywhere-h2020.eu/v1/bulkd_data?id_product=<PRODUCT_NAME>&date=<DATE>&end_date=<END_DATE>
Method	GET
Headers	X-API-KEY: <YOUR_API_KEY>
CURL instruction	`curl -X GET -H "X-API-KEY: <YOUR_API_KEY>" "https://rest.mhews.anywhere-h2020.eu/v1/bulk_data?id_product=<PRODUCT_NAME>&date=<DATE>&end_date=<END_DATE>" -o data.zip`
Expected result	A Zip file with the NetCDFs of the date interval.

3.2.2.2 Tests results

The following table depicts the results of each test for each partner:

Table 15: Summary table of the MH-EWS tests related to the Bulk Data service.

Test	CRAHI	CIMA	AIRBUS	RINAC	HYDS
Get files	OK	OK	OK	OK	OK
Wrong HTTP method	OK	OK	OK	OK	OK
No API-KEY	OK	OK	OK	OK	OK
Wrong API-KEY	OK	OK	OK	OK	OK
Get files of a non-existing product	OK	OK	OK	OK	OK
Get files of a non-contracted product	OK	OK	OK	OK	OK
Get files without a date	OK	OK	OK	OK	OK
Get files of a time interval	OK	OK	OK	OK	OK

3.2.2.3 Evaluation

Several errors raised when achieving Milestone 3.2. These errors were related with the HTTP error code returned by some operations and with dates format. All these errors were already fixed during the Milestone achivement and validated again. All the tests carried out to fulfill this document confirmed that all the modifications worked as expected, therefore no additional modifications nor improvements are foreseen in the bulk data services.

3.2.3 Geospatial data service

The Geospatial data service provides means of retrieving geospatial information such as maps based on the products served by the MH-EWS to complement the time series and the bulk data services.

3.2.3.1 Tests definition

The test related to the Geospatial data service are mainly focused on the MH-EWS capability to correctly retrieve and use data from the GeoServer, considering that the information must be available through the WMS and the WFS OGC services. The following sections describe the different tests performed. All the tests can be passed by typing the URL in a web browser.

The tests were carried out by several partners. Each of them had a specific API-KEY and product assigned, to test the functionalities with different kind of requests.

Get image

This test validates that the geospatial service returns the image of a specific product and date.

Table 16: Get image test.

URL	http://geodata.mhews.anywhere-h2020.eu/geoserver/mhews/wms?SERVICE=WMS&VERSION=1.1.1&REQUEST=GetMap&FORMAT=image/jpeg&TRANSPARENT=true&STYLES&LAYERS=mhews:<PRODUCT_NAME> &SRS=EPSG:4326&WIDTH=769&HEIGHT=433&BBOX=-180,-90,180,90&TIME=<DATE>
Expected result	An image with the desired product.

Non-existing product

This test validates that the geospatial service returns an error when requesting a non-existing product. This will always appear, since each product is defined as a layer in the GeoServer and it sends an error message when requesting a non-existing layer.

Table 17: Non-existing product test.

URL	http://geodata.mhews.anywhere-h2020.eu/geoserver/mhews/wms?SERVICE=WMS&VERSION=1.1.1&REQUEST=GetMap&FORMAT=image/jpeg&TRANSPARENT=true&STYLES&LAYERS=mhews:<DUMMY_PRODUCT_NAME> &SRS=EPSG:4326&WIDTH=769&HEIGHT=433&BBOX=-180,-90,180,90&TIME=<DATE>
Expected result	<?xml version="1.0" encoding="UTF-8" standalone="no"?> <!DOCTYPE ServiceExceptionReport SYSTEM "http://geoserver.hydsdev.net:80/geoserver/schemas/wms/1.1.1/WMS_exception_1_1_1.dtd"> <ServiceExceptionReport version="1.1.1" > <ServiceException code="LayerNotDefined" locator="layers"> Could not find layer mhews:<DUMMY_PRODUCT_NAME> </ServiceException></ServiceExceptionReport>

3.2.3.2 Tests results

The following table depicts the results of each test for each partner:

Table 18: Summary table of the MH-EWS tests related to the Geospatial Data service.

Test	CRAHI	CIMA	AIRBUS	RINAC	HYDS
Get image	OK	OK	OK	OK	OK
Non-existing product	OK	OK	OK	OK	OK

3.2.3.3 Evaluation

Several partners raised an error regarding the dates format when achieving the Milestone 3.2. This error was related to a misunderstanding with the date format of the different requests. The misunderstanding was clarified with the partners carrying out these tests, and no additional modifications were considered.

The tests passed to complete this document confirmed the service was working as expected and no additional modifications are expected.

3.2.4 FTP/FTPS service

The FTP/FTPS service is the service allows for data exchange through the File Transfer Protocol (FTP) or its extension FTP Secure (FTPS) that adds support for the Transport Layer Security (TLS) and the Secure Sockets Layer (SSL) cryptographic protocols.

3.2.4.1 Tests definition

The validation activities for the FTP/FTPS service focuses on the security (main access and data exchange) and the performance. The use of the FTPS protocol ensures safe data exchange, which is considered sufficient within the scope of this project.

Thus, the tests carried out cover both the server and client side checking the following aspects:

- Data transfer performance (i.e. throughput to guarantee the system availability under several requests, often called *stress test*).
- Access control (selective access to resources' writing).
- MH-EWS capability to access a remote FTP server to download and upload data.

The validation has been carried out using the FTP client named *FileZilla*. The following sections present the different tests carried out to validate different model providers that are uploading data, which will be integrated during the implementation phase.

Two partners passed the different tests defined in this section, to verify each of the functionalities offered by the service.

<u>FTP connection</u>

This test validates that the partner organization FMI can connect to the MH-EWS server to upload data.

Table 19: FTP connection test.

Description	Access to the FTP with the login and password.
Expected result	Connected to the FTP server.

Upload EFAS data

This test validates that the partner organization ECMWF can upload EFAS data to the MH-EWS server as an example of the data that is uploaded to the system.

Table 20: Upload EFAS data test.

Description	Access to the FTP with ECMWF login and password.
Expected result	Connected to the FTP server. A dummy file (text file or similar) can be uploaded.

3.2.4.2 Test results

The following table depicts the results of each test for each partner:

Table 21: Summary table of the MH-EWS tests related to the FTP/FTPS service.

Test	FMI	ECMWF
FTP connection	OK	
Upload EFAS data		OK

3.2.4.3 Evaluation

All these tests were successfully passed during the achievement of Milestone 3.2. Several partners have been using the service to provide and retrieve data without problems.

3.2.5 General functional tests evaluation

The different functional tests proposed and carried out confirmed that the MH-EWS is working operationally and as expected under the specifications.

3.3 Use cases tests

This section presents the different use cases tests carried out in the MH-EWS to assess its performance for the IT part. The aim of these tests is to validate the functionalities of the MH-EWS under certain conditions and provide indicators of the performance.

Deliverable 3.2 presented a detailed description of the use cases considered in the MH-EWS. In order to provide general indicators of the MH-EWS performance, three of these use cases have been selected as the most representative ones to assess the performance of the MH-EWS. These use cases cover the main actions that the users (*Actors* in the use case language) can perform, which are mentioned below:

1. Integrate external data into de MH-EWS.
2. Process data in the MH-EWS (run *encapsulated* algorithms).
3. Retrieve data from MH-EWS.

Different tests have been carried out on each use case. In addition, two different scenarios have been considered to provide a global view of the MH-EWS, using the same data set on each use case and testing the algorithms the MH-EWS is running for these data operationally. These scenarios are:

1. An algorithm that runs very frequently (every certain minutes), which has low computation time. In this case, FF-EWS algorithm has been run using OPERA radar data.
2. An algorithm with high computation time but that rarely run (e.g. monthly). In this case, WUR drought's algorithms have been run using EFAS seasonal data.

The following sections provide the performance assessment tests performed for each use case.

3.3.1 Integrate external data into the MH-EWS

The purpose of this use case is to incorporate new data to the MH-EWS. Thus, the user involved in this use case consists in a data provider that uploads data to the MH-EWS.

The Gateway is the MH-EWS module in charge of the acquisition of external data. Its goal is to acquire all the external information, convert it to the appropriate formats (if required), and insert it into the MH-EWS. The Gateway has several services (FTP, OGC SOS, OGC WFS, and others) with different interfaces that can be used by the data suppliers. A detailed description of the Gateway can be found in the Deliverable 3.2 (Section 2.5.3, page 38).

As explained in the Deliverable 3.2, the Gateway considers two different scenarios to integrate external data into the MH-EWS:

1. The MH-EWS retrieves data from an external data provider. Two different cases were considered to assess this scenario:
    a. Integration of RAQ model data, which is retrieved through a WSDL (Web Services Description Language) protocol. A description of RAQ models data can be found in the Deliverable 3.1 (Section 5.3, page 65).
    b. Integration of FMI Open data, which is retrieved through a WFS (Web Feature Service) protocol. A description of data provided by FMI can be found in the Deliverable 3.1 (Section 8.2, page 107).
2. The external data provider integrates the information into the MH-EWS. Two cases were considered for this scenario:
    a. Integration of EFAS Seasonal data, which is as an example of a high data integration. A description of EFAS Seasonal data can be found in the Deliverable 3.1 (Section 2.1, page 6).
    b. Integration of OPERA radar data, which is an example of frequently integrated data. A description of OPERA radar data can be found in the Deliverable 3.1 (Section 2.2.1, page 12).

The following sections present the performance evaluation of these four cases.

3.3.1.1 Integrate RAQ data

As previously explained, this test covers the integration of RAQ data, provided through WSDL standard protocol. Thus, the aim of this test is to assess RAQ data is properly acquired from the WSDL service and successfully integrated into the MH-EWS.

Test definition

This test validates that (i) the system correctly integrates the provided by the service, and that (ii) the download and processing time is less than the update frequency defined as the time between two consecutive simulations generated by the model.

Table 22: Integrate RAQ test definition.

Identifier	Test - 1
Description	Download, process and store RAQ data from the WSDL data provider service in the MH-EWS
Expected result	RAQ data is integrated in the MH-EWS in the appropriate internal format.

In general, the validation activities carried out to integrate external data in the MH-EWS focused on getting the proper output data in a reasonable time. Considering that the RAQ model is providing daily simulations (as described in Deliverable 3.2, Annex C) the following KPI can be defined:

Table 23: KPIs defined for Test - 1.

CODE	Description	Units	Value
KPI- 1	The acquisition and processing time must be lower than the update frequency of the products integrated by the Gateway process.	Time in minutes	Less than 24 hours (theoretically) Less than 1 hour (practically)

Tests results

Table 24 presents the results of the tests carried out during a one-week period for the above-mentioned test.

Table 24: Results of the Test - 1 (Integrate RAQ data).

Execution date	Result	Total time
2018-18-06 23:16	OK	16 min 36.9 s
2018-19-06 23:16	OK	16 min 50.9 s
2018-20-06 23:18	OK	18 min 5.7 s
2018-21-06 23:18	OK	18 min 3.6 s
2018-22-06 23:13	OK	13 min 13.5 s
2018-23-06 23:15	OK	15 min 25.9 s
2018-24-06 23:16	OK	16 min 57.5 s

Evaluation

As shown in the previous results, the time spent on downloading and integrating the RAQ data in the MH-EWS is within the limits defined by KPI- 1.

3.3.1.2 Integrate FMI Open Data

This test covers the integration of the Harmonie model data into the MH-EWS. This data is provided by FMI through the Web Feature Service (WFS) protocol. The aim of this test is to check that the MH-EWS is able to acquire and successfully integrate the Harmonie data from the WSDL service.

Test definition

The test validates that (i) the system correctly integrates the data provided by the service, and that (ii) the time spent to acquire and integrate the data is less than the update frequency (defined as the time elapsed between two consecutive simulations generated by the model).

Table 25: Integrate FMI Open Data test definition.

Identifier	Test - 2
Description	Download, process and store FMI Open Data from the WFS data provider service in the MH-EWS.
Expected result	Harmonie model data is integrated in the MH-EWS in the appropriate internal format.

The KPIs defined for this test are:

Table 26: KPIs defined for Test - 2.

CODE	Description	Units	Value
KPI- 2	The acquisition and processing must be lower than the update frequency of the products integrated by the Gateway process.	Time in minutes	Less than 6 hours (theoretically) Less than 3 hours (practically)

Test result

Table 27 presents the results of the tests carried out for the Test - 2.

Table 27: Results of the Test - 2 (Integrate FMI Open Data).

Execution date	Result	Total time
2018-25-06. 02:09	OK	1 h 19 min 32.4 s
2018-25-06. 03:27	OK	1 h 17 min 44.0 s
2018-25-06. 05:00	OK	1 h 32 min 25.8 s
2018-25-06. 06:20	OK	1 h 18 min 33.5 s
2018-25-06. 07:41	OK	1 h 19 min 8.3 s
2018-25-06. 09:18	OK	1 h 36 min 14.2 s
2018-25-06. 12:31	OK	2 h 37 min 58.5 s
2018-25-06. 14:26	OK	1 h 54 min 26.4 s
2018-25-06. 16:22	OK	1 h 54 min 46.9 s
2018-25-06. 18:21	OK	1 h 57 min 9.4 s
2018-25-06. 20:24	OK	2 h 2 min 28.6 s
2018-25-06. 22:22	OK	1 h 56 min 2.5 s

Evaluation

Considering the worst case found in the tests (2h 37m), the results are acceptable according to the KPI-2, hence no improvements are proposed.

3.3.1.3 Integrate EFAS Seasonal data

This test covers the integration of the EFAS Seasonal model data in the MH-EWS. These data set has been chosen due to its low update frequency and big size. EFAS Seasonal data is uploaded by the data provider in the FTP server, and from there it is integrated by the Gateway module in the MH-EWS.

Test definition

The assessment in this case is divided in two parts: (i) upload the data in the MH-EWS (done by the data provider), and (ii) integrate internally the data within the MH-EWS. The following Table 28 describes each of these tests in detail.

Table 28: Integrate EFAS Seasonal tests definition.

Identifier	Test - 3
Description	Upload data to the FTP server using ECMWF credentials.
Expected result	User connected to the FTP server and uploaded EFAS Seasonal data properly.

Identifier	Test - 4
Description	Integration of EFAS Seasonal data in the MH-EWS with the appropriate internal formats from the FTP server.
Expected result	EFAS Seasonal data is integrated in the MH-EWS.

The validation activities carried out to assess the integration of external data in the MH-EWS focused on integrating the data in a reasonable time period. Considering that the data is monthly updated, the following Table 29 presents the KPIs defined for this test.

Table 29: KPIs defined for Test - 3 and Test - 4.

CODE	Description	Units	Value
KPI- 3	The total time spent to upload and integrate the data must be lower than the update frequency of EFAS Seasonal model.	Time	Less than 1 month (theoretically) Less than 1 day (practically)

Tests results

The following Table 30 presents the results of uploading EFAS Seasonal data:

Table 30: Results of the Test - 3 (Uploading EFAS Seasonal data).

Execution date	File	Volume of data transferred (GB)	Result
2018-06-04 19:47:21	SEA2018060100_p00.tar	1.25	OK
2018-06-04 19:55:55	SEA2018060100_p01.tar	1.25	OK
2018-06-04 20:02:34	SEA2018060100_p02.tar	1.23	OK
2018-06-04 20:09:06	SEA2018060100_p03.tar	1.25	OK
2018-06-04 20:16:53	SEA2018060100_p04.tar	1.25	OK
2018-06-04 20:21:58	SEA2018060100_p05.tar	1.26	OK
2018-06-04 20:26:29	SEA2018060100_p06.tar	1.24	OK
2018-06-04 20:32:18	SEA2018060100_p07.tar	1.24	OK
2018-06-04 20:36:43	SEA2018060100_p08.tar	1.24	OK
2018-06-04 20:40:13	SEA2018060100_p09.tar	1.25	OK
2018-06-04 20:43:23	SEA2018060100_p10.tar	1.26	OK
2018-06-04 20:47:01	SEA2018060100_p11.tar	1.24	OK
2018-06-04 20:51:13	SEA2018060100_p12.tar	1.25	OK
2018-06-04 20:55:21	SEA2018060100_p13.tar	1.26	OK
2018-06-04 20:59:11	SEA2018060100_p14.tar	1.25	OK
2018-06-04 21:02:28	SEA2018060100_p15.tar	1.24	OK
2018-06-04 21:07:40	SEA2018060100_p16.tar	1.25	OK
2018-06-04 21:13:24	SEA2018060100_p17.tar	1.25	OK
2018-06-04 21:21:27	SEA2018060100_p18.tar	1.25	OK
2018-06-04 21:26:08	SEA2018060100_p19.tar	1.25	OK
2018-06-04 21:30:41	SEA2018060100_p20.tar	1.25	OK
2018-06-04 21:35:18	SEA2018060100_p21.tar	1.26	OK
2018-06-04 21:38:05	SEA2018060100_p22.tar	1.24	OK
2018-06-04 21:41:30	SEA2018060100_p23.tar	1.24	OK
2018-06-04 21:44:36	SEA2018060100_p24.tar	1.26	OK
2018-06-04 21:47:33	SEA2018060100_p25.tar	1.24	OK
2018-06-04 21:50:15	SEA2018060100_p26.tar	1.24	OK
2018-06-04 21:52:44	SEA2018060100_p27.tar	1.26	OK
2018-06-04 21:56:07	SEA2018060100_p28.tar	1.25	OK
2018-06-04 21:58:34	SEA2018060100_p29.tar	1.24	OK
2018-06-04 22:02:33	SEA2018060100_p30.tar	1.24	OK
2018-06-04 22:04:52	SEA2018060100_p31.tar	1.25	OK
2018-06-04 22:07:30	SEA2018060100_p32.tar	1.24	OK

Execution date	File	Volume of data transferred (GB)	Result
2018-06-04 22:10:16	SEA2018060100_p33.tar	1.24	OK
2018-06-04 22:12:59	SEA2018060100_p34.tar	1.25	OK
2018-06-04 22:15:29	SEA2018060100_p35.tar	1.25	OK
2018-06-04 22:17:49	SEA2018060100_p36.tar	1.25	OK
2018-06-04 22:20:43	SEA2018060100_p37.tar	1.25	OK
2018-06-04 22:23:20	SEA2018060100_p38.tar	1.24	OK
2018-06-04 22:25:38	SEA2018060100_p39.tar	1.24	OK
2018-06-04 22:27:59	SEA2018060100_p40.tar	1.26	OK
2018-06-04 22:30:05	SEA2018060100_p41.tar	1.25	OK
2018-06-04 22:32:39	SEA2018060100_p42.tar	1.24	OK
2018-06-04 22:35:53	SEA2018060100_p43.tar	1.25	OK
2018-06-04 22:38:13	SEA2018060100_p44.tar	1.24	OK
2018-06-04 22:40:42	SEA2018060100_p45.tar	1.25	OK
2018-06-04 22:43:13	SEA2018060100_p46.tar	1.25	OK
2018-06-04 22:45:38	SEA2018060100_p47.tar	1.25	OK
2018-06-04 22:47:57	SEA2018060100_p48.tar	1.25	OK
2018-06-04 22:50:31	SEA2018060100_p49.tar	1.25	OK

The following Table 31 presents the results of processing EFAS Seasonal data:

Table 31: Results of the Test - 4 (Integrating EFAS Seasonal data).

Execution date	Total time	Result
2018-18-06 23:16	5 h 44 min 58.2 s	OK

Evaluation

The total time spent by both tests is less than 9 hours (~3 hours and less than 6 hours respectively). This processing time complies the KPI- 3 definition, therefore no improvements are proposed.

3.3.1.4 Integrate OPERA radar data

This test covers the integration of OPERA radar data provided by FMI. OPERA radar network provides a new observation every 15 minutes, which can be considered a high refresh ratio. The test covers a similar scenario to the previous test, where data is uploaded in the MH-EWS FTP server and the Gateway module is in charge of its internal integration.

Test Definition

The aim of the test is to check that (i) users can upload data to the service, and that (ii) the uploading and processing time is less than the update frequency (defined as the time elapsed between two consecutive observations). The following Table 32 summarizes each of these tests.

Table 32: Integrate OPERA Data tests.

Identifier	Test - 5

Description	Upload data to the FTP server using FMI credentials.
Expected result	User connected to the FTP server and data is uploaded.

Identifier	Test - 6
Description	OPERA data is processed and integrated in the MH-EWS with the proper formats.
Expected result	OPERA data is integrated in the MH-EWS

Taking into account that OPERA observations are generated every 15 minutes, Table 33 details the list of KPIs defined for these tests:

Table 33: KPIs defined for Test - 5 and Test - 6.

CODE	Description	Units	Value
KPI- 4	The acquisition and processing time must be lower than the time between two consecutive observations.	Time in seconds	Less than 15 minutes

Test result

Table 34 presents the results of uploading the OPERA radar data set in the FTP server.

Table 34: Results of the Test - 5 (Uploading OPERA data).

File	Volume of data transferred (Kb)	Time (s)	Result
T_PAAH21_C_EUOC_20180624220000.hdf	930	1.1	OK
T_PAAH21_C_EUOC_20180624223000.hdf	921	1.0	OK
T_PAAH21_C_EUOC_20180624224500.hdf	917	0.9	OK
T_PAAH21_C_EUOC_20180624230000.hdf	905	1.0	OK
T_PAAH21_C_EUOC_20180624231500.hdf	907	1.0	OK
T_PAAH21_C_EUOC_20180624233000.hdf	902	1.0	OK
T_PAAH21_C_EUOC_20180624234500.hdf	892	1.1	OK
T_PAAH21_C_EUOC_20180625000000.hdf	900	1.0	OK
T_PAAH21_C_EUOC_20180625001500.hdf	908	1.0	OK
T_PAAH21_C_EUOC_20180625004500.hdf	894	1.0	OK
T_PAAH21_C_EUOC_20180625010000.hdf	906	1.0	OK
T_PAAH21_C_EUOC_20180625011500.hdf	891	1.2	OK
T_PAAH21_C_EUOC_20180625014500.hdf	863	1.0	OK
T_PAAH21_C_EUOC_20180625020000.hdf	855	1.1	OK
T_PAAH21_C_EUOC_20180625021500.hdf	841	1.0	OK
T_PAAH21_C_EUOC_20180625023000.hdf	842	1.0	OK
T_PAAH21_C_EUOC_20180625024500.hdf	839	1.1	OK
T_PAAH21_C_EUOC_20180625030000.hdf	844	1.0	OK
T_PAAH21_C_EUOC_20180625031500.hdf	819	1.0	OK
T_PAAH21_C_EUOC_20180625033000.hdf	812	1.0	OK
T_PAAH21_C_EUOC_20180625034500.hdf	781	1.1	OK
T_PAAH21_C_EUOC_20180624220000.hdf	930	1.1	OK
T_PAAH21_C_EUOC_20180624223000.hdf	921	1.0	OK

The following Table 35 summarizes the time spent to integrate several the OPERA radar data observations during a 6-hour period.

Table 35: Results of Test - 6 (Integrating OPERA data).

Execution date	Total time (s)	Result
2018-25-06 00:00	8.8	OK
2018-25-06 00:15	8.5	OK
2018-25-06 00:30	5.5	OK
2018-25-06 00:59	3.1	OK
2018-25-06 01:14	3.2	OK
2018-25-06 01:29	4.1	OK
2018-25-06 01:59	2.2	OK
2018-25-06 02:15	9.4	OK
2018-25-06 02:29	2.3	OK
2018-25-06 02:44	2.8	OK
2018-25-06 03:00	4.7	OK
2018-25-06 03:15	8.7	OK
2018-25-06 03:29	2.3	OK
2018-25-06 03:46	3.1	OK
2018-25-06 04:00	10.0	OK
2018-25-06 04:30	4.8	OK
2018-25-06 04:44	3.0	OK
2018-25-06 05:00	7.8	OK
2018-25-06 05:15	7.5	OK
2018-25-06 05:46	2.4	OK
2018-25-06 06:01	4.3	OK
2018-25-06 06:15	7.8	OK

Evaluation

Considering the worst case found in the two tests (1.22 s for the Upload time and 10.07s for processing the data) the results fulfill the KPI- 4 hence no improvements are proposed.

3.3.2 Processing data

The purpose of this use case is to check that the encapsulated models run properly within the MH-EWS and their outputs are integrated into the system.

As previously mentioned in section 3.3, two different scenarios were considered for this use case: (i) to run FF-EWS model with the OPERA data, and (ii) to run WUR drought's model with EFAS seasonal data. The following sections present in detail the tests carried out to assess the use case under these two scenarios.

3.3.2.1 FF-EWS model

This test covers the process performed within the MH-EWS every time new OPERA radar data set is integrated into the system. Thus, the test can be considered as the continuation of the test presented in section 3.3.1.4.

Test definition

The test aims to validate that the model generates the appropriate outputs in a reasonable time and integrates them in the MH-EWS. The following Table 36 summarizes this test.

Table 36: FF-EWS model test.

Identifier	Test - 7
Description	Run the FF-EWS model to generate a new radar nowcasting using the OPERA data observations.
Expected result	A new radar nowcasting is integrated in the MH-EWS.

The update frequency of the FF-EWS model is defined by the update frequency of the input data. Considering OPERA data is received every 15 minutes, the KPIs for this test are defined in the following Table 37:

Table 37: KPIs defined for Test - 7.

CODE	Description	Units	Value
KPI- 5	Time spent to generate a new nowcasting and to integrate it must be lower than input product's update frequency.	Time in minutes	Less than 15 minutes

Test result

Preliminary tests showed the total processing time exceeded the 15-minutes threshold defined by KPI- 5. After some research, it was found that the bottleneck was on the data integration process (~10 minutes) rather than in the model calculation itself (~6 minutes), so a modification in the MH-EWS was needed to achieve the thresholds defined in this test.

Improvement

The MH-EWS stores internally the raster data in NetCDF format. Although this format is supported by the GeoServer (which is part of the Data supply module of the MH-EWS), the updates and requests of raster maps stored in this format in the GeoServer offer a poor performance.

This implies that several updates are carried out in GeoServer every time the FF-EWS output data is integrated in the MH-EWS. In order to increase the efficiency, a modification in the MH-EWS data integration process was introduced to store data in the system in both NetCDF and GeoTiff formats. Thus, GeoTiff files are used by GeoServer, reducing its response time when updating and requesting data.

Tests results after improvements

Table 38 presents the results of the tests carried out after applying the modifications previously mentioned:

Table 38: Results of Test - 7 (Processing FF-EWS model with OPERA radar data).

Execution date	Result	Total time
2018-25-06 11:39	OK	8m 53.9s
2018-25-06 11:54	OK	9m 14.8s
2018-25-06 12:11	OK	11m 46.5s
2018-25-06 12:25	OK	10m 46.7s
2018-25-06 12:53	OK	9m 17.1s
2018-25-06 13:10	OK	11m 7. 2s
2018-25-06 13:25	OK	10m 32.9s
2018-25-06 13:39	OK	10m 8.7s
2018-25-06 11:39	OK	8m 53.9s

Evaluation

Following the results of the previous table, the total time obtained met the KPI- 5.

3.3.2.2 WUR drought's model

This test covers the process performed within the MH-EWS every time new EFAS Seasonal data is integrated in the system. Thus, the test can be considered as the continuation of the test presented in section 3.3.1.3.

Test definition

As in the previous test, the aim of the assessment of the Processing data use case is to check that the model generates its outputs in a certain time and that they are stored in the MH-EWS. The following Table 39 summarizes this test.

Table 39: WUR drought's model test.

Identifier	Test - 8
Description	Run WUR drought's model to generate a new drought forecasting using EFAS Seasonal data and store these forecast in the MH-EWS.
Expected result	A new drought forecast is available in the MH-EWS.

EFAS generates a new seasonal simulation every month. Taking this into account, the KPI defined for this test is indicated in the following Table 40:

Table 40: KPIs defined for Test - 8.

CODE	Description	Units	Value
KPI- 6	Time spent to generate a new forecasting and storing it must be lower than input products' update frequency.	Time	Less than 1 month (theoretically) Less than 1 day (practically)

Test result

WUR drought's model can be considered as a set programs that generate several products. All these programs use the same input data. The following Table 41 presents the results to generate these products.

Table 41: Results of Test - 8 (Processing WUR drought's model with EFAS Seasonal data).

Product	Execution date	Result	Total time
Standard Precipitation Index (SPI)	2018-13-06 09:33	OK	1h 16 min 59 s
Standard Groundwater Index (SGI)	2018-13-06 08:10	OK	46 min 5.9 s
Area Groundwater Index	2018-12-06 12:40	OK	1h 40 min 58.2 s
Area Precipitation Index	2018-27-06 08:10	OK	28 min 59.5 s
Soil Moisture Max Drought Start	2018-12-06 23:44	OK	24 min 54.7 s
Runoff Max Drought Start	2018-12-06 23:19	OK	27 min 57.4 s
Groundwater Max Drought END	2018-12-06 22:25	OK	1h 33 min 14.6 s
Runoff Drought Duration	2018-12-06 17:13	OK	1h 16 min 4.8 s
Precipitation Drought	2018-13-06 11:37	OK	29 min 28.8 s

Evaluation

The results presented in the previous section met KPI- 6 definition, confirming that it is working properly and no additional tasks are needed.

3.3.3 External user obtains models information from the MH-EWS

The purpose of this use case is to provide data generated by the different models of the system to the final users through different interfaces.

Several specific scenarios have been considered to assess this use case, namely:

- Request EFAS Seasonal data using the Bulk data service.
- Obtain OPERA radar data using the Bulk data service as well.
- Obtain OPERA geospatial data through the GeoSpatial data service.
- Obtain the products catalogue, retrieving the metadata of the products available in the MH-EWS.

The following sections present the performance assessment of this use case considering the above mentioned scenarios.

3.3.3.1 Obtain EFAS Seasonal data

This test covers the request of the EFAS Seasonal data from the MH-EWS through the Bulk data service. The EFAS Seasonal data consist of 6 products (discharge, evaporation, precipitation, total runoff, soil moisture in the upper layer and storage in the upper layer). The EFAS Seasonal data size is very big, and its data request can be considered as the final step of the EFAS Seasonal data integration.

Test Definition

The aim of this test is to assess the performance of the MH-EWS when providing EFAS Seasonal data. The following Table 42 presents this test.

Table 42: Obtain EFAS Seasonal data test.

Identifier	Test - 9
URL	https://rest.mhews.anywhere-h2020.eu/v1/bulkd_data?id_product=<PRODUCT_NAME>&date=<DATE>&end_date=<DATE>
Method	GET
Headers	X-API-KEY: XXXXX
cURL instruction	`curl -X GET -H "X-API-KEY: XXXXXXXX" "https://rest.mhews.anywhere-h2020.eu/v1/bulk_data?id_product=<PRODUCT_NAME>&date=<DATE>&end_date=<DATE>" -o data.zip`
Expected result	A Zip file with the NetCDFs of the given interval

According to the test, Table 43 summarizes the KPI identified for this test:

Table 43: KPIs defined for Test - 9.

CODE	Description	Units	Value
KPI- 7	Time spent to acquire the data must be less than a reasonable period of time.	Time in seconds	Less than 10 seconds

Test Results

Table 44 summarizes the results of the tests performed for each EFAS Seasonal product.

Table 44: Results of Test - 9 (Obtaining EFAS Seasonal data).

Product	Result	Total time (s)	Size (Kb)
Discharge	OK	2.6	7537
Evaporation	OK	2.2	603
Precipitation	OK	2.3	692
Total runoff	OK	2.6	7378
Soil moisture upper	OK	2.6	7039
Storage upper	OK	3.8	5067

Evaluation

Several factors must be considered when analyzing these results (the amount of requested data, the server bandwidth, the client bandwidth, the server load, etc.). Although these issues, the response time of the requests is very low compared with the KPIs defined for the test, so no additional modifications or improvements are expected.

3.3.3.2 Obtain OPERA data

This test covers the request of the OPERA radar data from the Bulk data service, which has a higher time resolution but with a lower size than the EFAS Seasonal data (whose data request test was presented in the previous section).

Test definition

As for the previous test, the aim of this test is to assess the performance of the MH-EWS when providing OPERA radar data. The following Table 45 presents this test.

Table 45: Obtain OPERA data test.

Identifier	Test - 10
URL	https://rest.mhews.anywhere-h2020.eu/v1/bulkd_data?id_product=opera_rain_rate&date=<DATE>&end_date=<DATE>
Method	GET
Headers	X-API-KEY: XXXXX
cURL instruction	`curl -X GET -H "X-API-KEY: XXXXXXXX" "https://rest.mhews.anywhere-h2020.eu/v1/bulk_data?id_product=opera_rain_rate&date=<DATE>&end_date=<DATE>" -o data.zip`
Expected result	A Zip file with the NetCDFs of the date interval.

The tests carried out for the OPERA data consisted on requesting 2 hours of radar data set. As in the previous test, KPI- 7 will be used as a reference to assess the use case performance.

Test results

Table 46 summarizes the results of this test:

Table 46: Results of the Table 10 (Obtain OPERA data).

Product name	Result	Total time (s)	Size (Kb)
Instantaneous Surface Rain rate	OK	5.4	5067

Evaluation

This test is also affected by the same factors mentioned in the previous test (the period of time demanded, the server bandwidth, the client bandwidth, the server load). Even though, the results fulfill the requirements defined by KPI- 7, therefore no additional modifications nor improvements are considered at this point.

3.3.3.3 Obtain OPERA geospatial data

This test covers the maps request of products available in the MH-EWS using the Geospatial data. The Geospatial data service provides geospatial information through WMS protocol defined by OGC.

Test definition

This test will consist in requesting maps using WMS protocol through the Geospatial data service. The following Table 47 summarizes this test.

Table 47: Obtain OPERA geospatial data test.

Identifier	Test - 11
URL	`http://geodata.anywhere-h2020.eu/geoserver/mhews/wms?SERVICE=WMS&VERSION=1.1.1&REQUEST=GetMap&FORMAT=image/jpeg&TRANSPARENT=true&STYLES&LAYERS=mhews:<PRODUCT_NAME> &SRS=EPSG:4326&WIDTH=769&HEIGHT=433&BBOX=-180,-90,180,90&TIME=<DATE>`
Expected result	An image with the desired product retrieved in a reasonable time.

In this case, requested products will be those generated by the FF-EWS model using OPERA radar data (river warning, rain warning, 15-minutes rain accumulation, 1-hour rain accumulation and 24-hours rain accumulation). All the available maps between 2018-06-27 13:15 and 2018-06-28 16:00 have been requested for each of these products.

The KPIs defined for this test are listed in the following Table 48.

Table 48: KPIs defined for Test - 11.

CODE	Description	Unit	Value
KPI- 8	Retrieve the image map.	Time in seconds	Less than 5 seconds

Test result

Table 49 summarizes the results of the different tests performed, considering the conditions mentioned above (several requests within a time interval for each product).

Table 49: Results of Test - 11 (Obtain OPERA geospatial data).

Product	Min (s)	Max (s)	Average (s)
River warning	0.3	3.3	1.8
Rain warning	0.2	0.4	0.3
Rain accumulation 24h	0.2	0.4	0.3
Rain accumulation 1h	0.3	0.6	0.4
Rain accumulation 15min	0.2	0.6	0.4

Evaluation

All these tests were successfully passed. The Geospatial data service has been working operationally for several months, and no incidents have been reported. The obtained response times are under the thresholds defined by KPI- 8, so no additional measures are considered.

3.3.3.4 Obtain products catalogue

This test covers the performance assessment of the Products' catalogue service. This service provides meta-information about the products available in the MH-EWS and it is a service intended for machine-to-machine interaction that consists of a secure REST API. The design as well as the definition of the service interface is available in the Deliverable 3.2 (Section 2.5.4.2, page 43 and Annex A, page 108, respectively).

Test definition

The aim of this test is to prove that the Products' catalogue service can provide the products information for a given user. The following Table 50 defines this test.

Table 50: Obtain products catalogue test.

<table>
<tr><td>Identifier</td><td>Test - 12</td></tr>
<tr><td>URL</td><td>https://rest.mhews.anywhere-h2020.eu/v1/prod_catalogue</td></tr>
<tr><td>Method</td><td>GET</td></tr>
<tr><td>Headers</td><td>X-API-KEY: <YOUR_API_KEY></td></tr>
<tr><td>Expected result</td><td><pre>
[
    {
        "name":"ifs_hres_10m_u_wind_speed",
        "description":"10 meters U wind speed",
        "unit":"m/s",
        "id_product":"10m_u_wind_speed",
        "time_step":"10800",
        "update_frequency":"604800",
        "first_data":"2017-12-05 18:00:00",
        "last_data":"2017-12-24 00:00:00",
        "data_type":"raster",
        "bounding_box":"[-27,33,45,73.5]",
        "resolution":"[0.1,0.1]",
        "available_formats":"raster"
    },
    {
        "name":"ifs_hres_10m_v_wind_speed",
        "description":"10 meters V wind speed",
        "unit":"m/s",
        "id_product":"10m_v_wind_speed",
        "time_step":"10800",
        "update_frequency":"604800",
        "first_data":"2017-12-05 18:00:00",
        "last_data":"2017-12-24 00:00:00",
        "data_type":"raster",
        "bounding_box":"[-27,33,45,73.5]",
        "resolution":"[0.1,0.1]",
        "available_formats":"raster"
    },
    {
        "name":"ifs_hres_2m_temperature",
        "description":"2 meters temperature",
        "unit":"°K",
        "id_product":"2m_temperature",
        "time_step":"10800",
        "update_frequency":"604800",
        "first_data":"2017-12-05 18:00:00",
        "last_data":"2017-12-24 00:00:00",
        "data_type":"raster",
        "bounding_box":"[-27,33,45,73.5]",
        "resolution":"[0.1,0.1]",
        "available_formats":"raster"
    },
    {
        "name":"ifs_hres_dew_point_temperature",
        "description":"Dew point temperature",
        "unit":"°K",
        "id_product":"dew_point_temperature",
        "time_step":"10800",
        "update_frequency":"604800",
        "first_data":"2017-12-05 18:00:00",
        "last_data":"2017-12-24 00:00:00",
        "data_type":"raster",
        "bounding_box":"[-27,33,45,73.5]",
</pre></td></tr>
</table>

```
        "resolution":"[0.1,0.1]",
        "available_formats":"raster"
    },
    {
        "name":"ifs_hres_precipitation",
        "description":"Precipitation",
        "unit":"mm",
        "id_product":"precipitation",
        "time_step":"86400",
        "update_frequency":"604800",
        "first_data":"2017-12-05 18:00:00",
        "last_data":"2017-12-24 00:00:00",
        "data_type":"raster",
        "bounding_box":"[-27,33,45,73.5]",
        "resolution":"[0.1,0.1]",
        "available_formats":"raster"
    }
]
```

This test has been executed with different users with a different number of products assigned.

The KPIs identified for this test are listed in the following Table 51:

Table 51: KPIs defined for Test - 12.

CODE	Description	Units	Value
KPI- 9	The response time must be lower than the defined valued.	Time in seconds	10 seconds

It is important to remark that for response time for the service must be independent of the number of available products for a given user.

Tests results

Preliminary results showed that some requests took more than one minute in some cases. This mainly happened when users have a large set of products. In order to avoid this situation a modification of the Products' Catalogue service was performed to improve its performance.

Improvement

After an analysis of the procedure of how the service collects product's information from the Internal database (using its API REST), it was concluded that the use of the internal REST API operations is not suitable in this case. Thus, Product's catalogue operation was redefined to make direct requests to the database surpassing the API REST.

Tests results after performance assessment improvements

The following Table 52 presents the results of the different requests performed to the Products' Catalogue service:

Table 52: Results of Test - 12 (Obtain products catalogue).

Description	Result	Time elapsed (s)
Assigned products 1	OK	0.26
Assigned products 2	OK	0.21

Description	Result	Time elapsed (s)
Assigned products 4	OK	0.24
Assigned products 13	OK	0.21
Assigned products 26	OK	0.27
Assigned products 68	OK	0.21
Assigned products 70	OK	0.24
Assigned products 296	OK	0.24
Assigned products 869	OK	0.25

Evaluation

According to these results, the modifications made in the service reduced the response time, meeting the thresholds defined by KPI- 9.

4 Algorithm/Product performance assessment by Pilot Site

This Chapter presents the performance assessment of the MH-EWS by providing the methodologies applied for the validation of the products and thus the algorithms there-in integrated and/or encapsulated.

The chapter presents, product by product, the methodology applied in the Pilot Site (i.e. the activities that will be carried out during the operational demonstration phase for products validation and/or tuning).

The chapter presents for each product under evaluation:

- The product ID in brackets (in case of a set, the dash is used),
- the scenario,
- the data set,
- the skill scores, and
- (if already available) the experience already done on the Pilot Sites.

4.1 ECMWF - Performance assessment of the Meteorological Forecasts and Nowcasts products [PRD-183, 187]

Table 53: Performance assessment summary for Meteorological Forecasts and Nowcasts products (ECMWF).

Pilot Sites of implementation	All the Pilot Sites, depending on the variable analysed and the possible events (extreme wind speed, low/high temperatures, precipitation, snow).
Description	Different variables provided from the ECMWF's Integrated Forecasting System (IFS), from both, high-resolution deterministic forecasts and ensemble forecast system are to be evaluated.
Method of evaluation	The method will consist in calculating different verification scores for probabilistic forecast (e.g. CRPS, ROC, Relative economic value). For the deterministic variables, scores such as false alarm rate or probability of detection will be evaluated. The new results will be used alongside ECMWF's existing forecast evaluation in the final evaluation.
References or other data used for validation	The verification of the meteorological parameters will be performed using in-situ meteorological measurements. Preferably these will be 3-hourly observations of manned and automatic SYNOP type stations in Europe. In case of not having access to observed data from within the pilot area ECMWF will seek to source representative data from other sources e.g. METAR observations.
Skill scores	The skill scores that will be computed depend on the number of samples of each parameter that we have during the study period. They will be based on the skill scores used in Richardson et al. (2012). In case of having a very reduced number of events, it will just be provided the hit rates, false alarms and misses.
Examples	Samples of how we can evaluate the performance of the products that can be found in Richardson et al. (2012).

4.1.1 Description of the performance assessment scenario

Different variables are provided from the ECMWF's Integrated Forecasting System (IFS), from both, high-resolution deterministic forecasts and ensemble forecast system:

- The Daily total precipitation probability is a variable of ECMWF's Integrated Forecasting System (IFS). It measures the 24-hour accumulated precipitation (rain and snow) exceeding 1, 5, 10 and 20 mm (in %). The probabilities are based on the number of forecast members which meet the criteria (each member is assigned an equal probability of 1/50). Total precipitation from high-resolution forecast will be analysed as well.

- The Extreme Forecast Index (EFI) ranks the departures between the Ensemble Prediction System forecasts and the model climate to identify unusual meteorological situations. EFI is applied on a daily basis to the following variables: temperature, wind speed, wind gust, Convective Available Potential Energy (CAPE), snowfall and total precipitation. Experience suggests that EFI values of 0.5 – 0.8 (irrespective of sign) can be generally regarded as signifying that "unusual" weather is likely and values above 0.8 as usually signifying that "very unusual" or extreme weather is likely.

- Wind gust probability above a given threshold. It is a variable of ECMWF's Integrated Forecasting System (IFS). It measures the maximum wind gusts exceeding 15, 20 and 25 m/s (in %) within a 120-hour time window. The probabilities are based on the number of forecast members which meet the criteria (each member is assigned an equal probability of 1/50). Note that the 10m wind gust is a post-processed product, computed as the sum of three terms: the instantaneous 10m wind speed, turbulent gustiness in the boundary layer and gustiness in convective situations.

It will be a quantitative verification based mainly in previous researches, as the one in Richardson et al. (2012) and it will be developed in the Pilot Sites by ECMWF during the period from June 2018 to May 2019.

4.1.2 Description of the input data and reference datasets

We are using the ensemble forecast from IFS with 1 control and 50 perturbed forecasts and the high-resolution forecast from ECMWF. For the verification of the ECMWF meteorological variables we will use the 0000 UTC base time and evaluated up to a lead time of 168 hours. In this ensemble configuration, the spatial resolution is approximately 18 km and 9 km in the high-resolution forecast, while the temporal resolution of the output available for the project is 3-hourly from 0 to 144 hours, and 6 hourly from144 hours onward. For the verification, we split the 7 days into 24-hour periods.

Observed data from within the Pilot Site will be used. If the temporal resolution (ideally 3 hourly) or spatial coverage of this data is limited then the ECMWF will seek to use other data by considering first the 3-hourly observations of present weather from automatic or manned SYNOP stations in Europe. However, METAR observations or other type of observations will be considered if we have not SYNOP stations close to the Pilot Site. Only 3-hourly verification will be possible as a result of the time resolution of the product.

All the data from meteorological forecasts products will be available for the evaluation of the Pilot Sites during the study period. The final result of the verification depends on the availability of the observations and other resources.

4.1.3 Description of the evaluation skill scores

The results of the verification will be compared with other previous studies in the Pilot Site (or similar areas) and ongoing forecast monitoring at the ECMWF.

It will be a quantitative evaluation, comparing the different verification scores, when they are available.

4.1.4 Experience and examples on the Pilot Sites

The results in the Pilot Sites will be evaluated as well by means of different case studies, showing how much the forecast agrees the observations at different lead times.

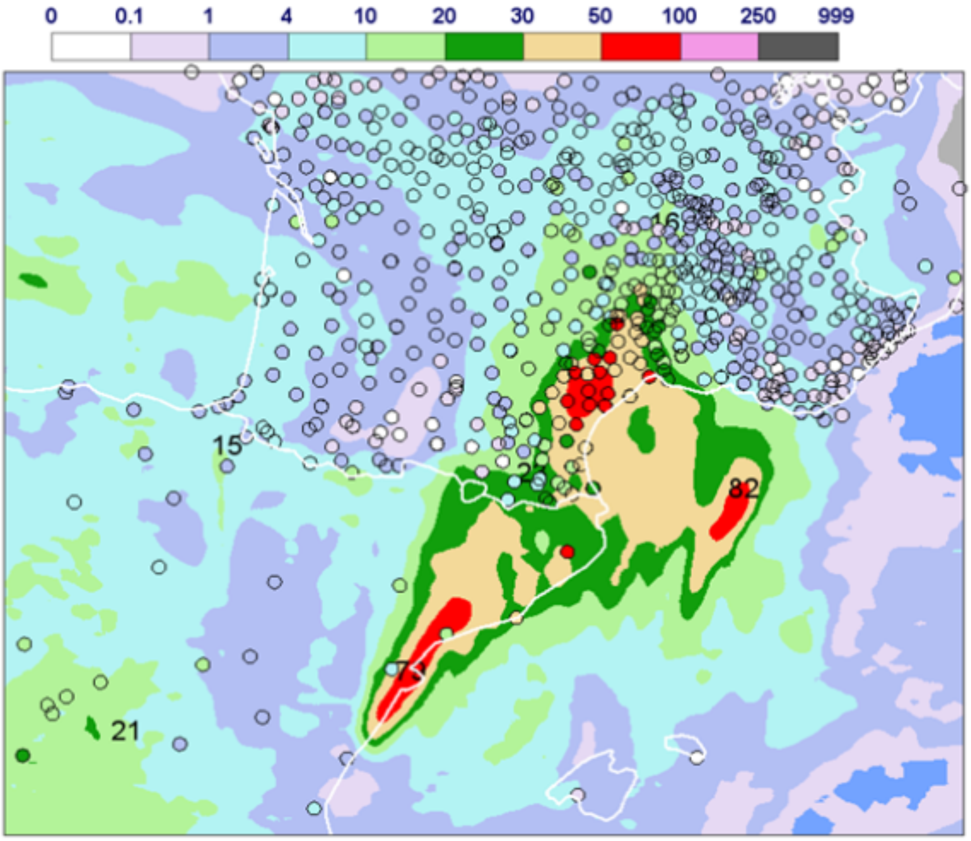


Figure 2: Precipitation totals for 48 hours from high-resolution ECMWF deterministic forecast at lead time 18-66h, compared with available observations in Catalonia and South France.

4.2 ECMWF – Performance assessment of probability of precipitation type [PRD 182-193] and most probable precipitation [PRD 194]

Table 54: Performance assessment summary for the products probability of precipitation type (PRD 182-193) and most probable precipitation (PRD 194).

Pilot Sites of implementation	Finland, Switzerland, Catalonia.
Description	Based on the IFS atmospheric model ensemble with 1 control and 50 perturbed forecasts, the probability of precipitation type (PRD 182-193) is calculated from the precipitation type variable that has 6 different precipitation categories: rain, freezing rain, snow, wet snow, sleet and ice pellets. The precipitation rate variable was used to classify the probabilities in different precipitation intensities. The most probable precipitation type describes which type of precipitation (rain, sleet, wet snow, snow, freezing rain or ice pellets) is most probable whenever the probability of some precipitation is >50%.
Method of evaluation	The method will consist of calculating different verification scores suitable for probabilistic forecasts (e.g. CRPS, ROC, ...) for the probability of precipitation type product. For the most probable precipitation type, the method will consist of calculating different verification scores suitable for deterministic forecasts (e.g. probability of detection, false alarm rate, performance diagram). Both verifications will be based on the methodology applied in the article: “Improving Predictions of Precipitation Type at the Surface: Description and Verification of Two New Products from the ECMWF Ensemble" (Gascón et al., 2017)
References or other data used for validation	Preferably the verification of both products will be performed using 3-hourly observations of present weather from manned SYNOP stations in Pilot Sites. In such data is not available automatic SYNOP stations, METAR observations or other observation that provides precipitation type information for the Pilot Sites will be considered.
Skill scores	The skill scores computed will depend on the number of samples of each precipitation type made during the study period. They will be based on the official verification scores used in the ECMWF, such as ROC, reliability or relative economic value for probabilistic forecasts. In case of having a very reduced number of cases, we will just provide the hit rates, false alarms and misses. For the most probable precipitation type we will consider the use of performance diagrams (to be able to evaluate several precipitation types with different skill scores in the same plot) or the symmetric extremal dependency index (SEDI) that can be useful to evaluate extreme events such as snow or freezing rain.
Examples	Samples of how we can evaluate the performance of the product in case studies can be found in Gascón et al. (2017) and in Gascón et al. (2018).

4.2.1 Description of the performance assessment scenario

The instantaneous probability of precipitation type (%), shown for a given Pilot Site, depicts the temporal evolution of precipitation type probabilities (rain, sleet, wet snow,

snow, freezing rain and ice pellets) for a specific location in bar chart format, to help the user make better decisions regarding particular events. The probability of precipitation type is displayed combined with the instantaneous total precipitation rate (another IFS variable) to provide, for example, an indication of potential freezing rain events and their likely severity, heavy snowfalls, etc. In this aspect, each precipitation type is also divided into 3 different categories depending on the precipitation rate, from one minimum value to 0.2 mm/h (low intensity), from 0.2 to 1 mm/h (medium intensity) and greater to 1 mm/h (high intensity).

Ideally, this product is plotted as a meteogram, showing the probabilities of each precipitation type for a specific location. The probabilities of each precipitation type depending on three precipitation rate categories are provided separately in grib format.

The verification of the probability of precipitation type product will be quantitative and probabilistic and it will be developed for the Pilot Sites by ECMWF during the period from June 2018 to May 2019.

The most probable precipitation type product is a secondary product from the probability of precipitation type algorithm that calculate which of the six precipitation types (rain, sleet, wet snow, snow, freezing rain and ice pellets) is most probable whenever the probability of some precipitation is >50%. Also, the most probable precipitation type is classified in three different ranges of probabilities: up to 50%, from 50 to 70% and higher than 70%. In order to give more useful information to the users, with probability of precipitation less than 50% the product stablish other two categories (grey colours) when the probability of any type of precipitation is between 10-30% and 30-50% giving, in this case, only information about the probability of occurrence of precipitation/no precipitation.

The verification of the most probable precipitation type will be quantitative but as categorical variables (occurrence or not of a specific type of precipitation) and it will be developed for the Pilot Sites by the ECMWF during the period from June 2018 to May 2019.

4.2.2 Description of the input data and reference datasets

We are using the precipitation type variable from ensemble forecast with 1 control and 50 perturbed forecasts (the IFS ensemble forecast). For the verification of the ECMWF probability of precipitation type product we will use the 0000 UTC base time and evaluated up to a lead time of 168 hours. In this ENS configuration, the spatial resolution is approximately 18 km while the temporal resolution of the output available for the project is 3-hourly from 0 to 144 hours, and 6 hourly from 144 hours onward. The products will be evaluated without taking into account the intensity of each precipitation type (precipitation rate variable).

Ideally 3-hourly observations of present weather from manned SYNOP stations in the Pilot Sites will be used, since the present weather parameter is not very accurate in automatic stations, especially for mixed precipitation (Elmore et al., 2015). However, automatic SYNOP stations, METAR observations or other observation that provides precipitation type information will be considered if there is no manned SYNOP station in the Pilot Site. Only 3-hourly verification will be possible as a result of the absence of more frequent SYNOP manual observations and the time resolution of the product. The original weather reports will be classified into one of five different categories: rain, snow, wet snow, freezing rain and ice pellets following the same methodology than Gascón et al., (2017). Wet snow will not be considered separately because of the lack of direct observations for its verification; instead, the wet snow forecasts will be classified as snow. The most probable precipitation type will be evaluated as "deterministic" forecast, since we will not consider the probabilities, if not the occurrence/non-occurrence of the event.

All the data from the precipitation type products will be available for the evaluation of the Pilot Sites during the study period. The final result of the verification depends on the availability of the observations.

4.2.3 Description of the evaluation skill scores

The results of the verification will be compared with other studies such as Gascon et al. (2017) and Fehlmann et al. (2018) to see the differences between different study areas, depending on the topography or the latitude.

It will be a quantitative evaluation, comparing the different verification scores, when they are available.

4.2.4 Experience and examples on the Pilot Sites

The results in the Pilot Sites will be evaluated as well by different case studies, showing how much the forecast agrees the observations at different lead times. Some examples can be found in Gascón et al., (2017) and Gascón et al., (2018).

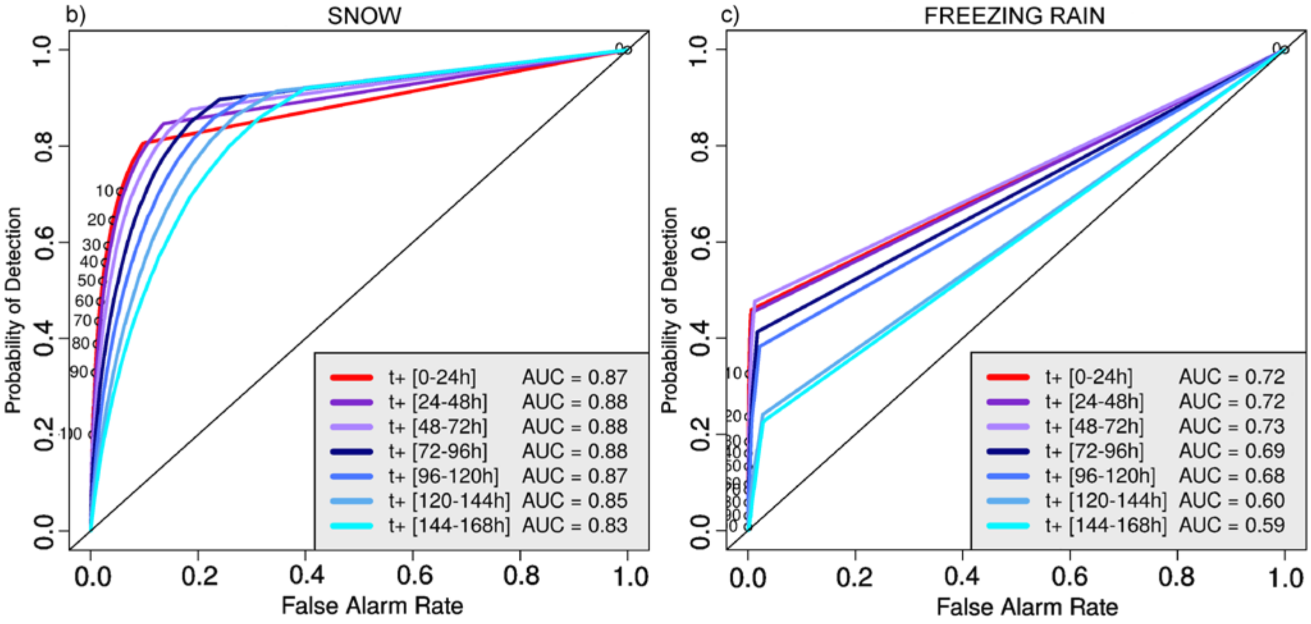


Figure 3: ROC curves at different lead times, up to day 7, for snow and freezing rain in Europe from the product probability of precipitation type. The curves are

the plots of hit rate vs false alarm rate for each decision threshold (2% interval used). Labels, at 10% intervals, are shown for the day-1 forecasts only (in red). The black line represents no skill. The AUC for each lead time is shown in a grey box.

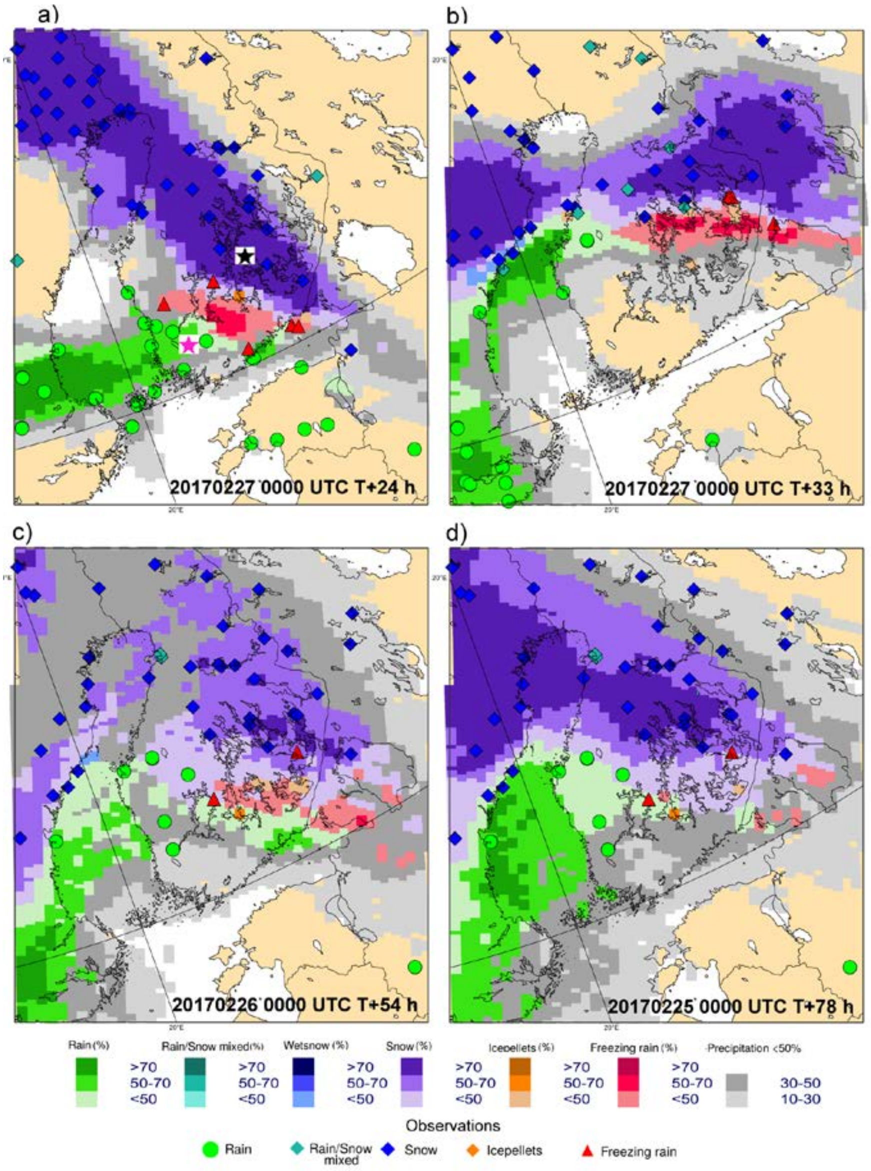


Figure 4: A significant freezing rain event occurred across southeast Finland on 28 Feb 2017. Map products show the most probable precipitation type, valid 28 Feb, at different base times. The observed precipitation types from SYNOP reports at the same times are plotted as symbols (dry is not shown). (a) The 24-h lead-time forecast for valid time 0000 UTC 28 Feb (pink star is the sounding site at Jokioinen and black star is Mikkeli). (b) The 33-h forecast for valid time 0900 UTC 28 Feb. (c) The 54-h forecast for valid time 0600 UTC 28 Feb. (d) The 78-h forecast for valid time 0600 UTC 28 Feb.

4.3 CIMA - Performance assessment of the Nowcasting Rainfall field [PRD 40]

Table 55: Performance assessment summary for algorithm PhaSt [PRD 40]

Pilot Sites of implementation	Liguria.
Description	PhaSt (Metta et al., 2009) is an ensemble radar-based rainfall nowcasting technique based on the extrapolation of rainfall observations by a diffusive process in the Fourier space. The extrapolation of radar observations is done maintaining the power spectral amplitudes constant and the Fourier phases evolve by an Ornstein-Uhlenbeck stochastic process (Rebora and Silvestro, 2012, Poletti et al, 2017) following a Langevin-type model, whose random component is sampled to characterize the uncertainty of the technique.
Method of evaluation	The assessment performance will be conducted in two different approach. One more qualitative verifying if the real observed rainfall event hits the area predicted by the model, and one quantitative using the PhaSt output in the hydrologic model (Flood-PROOFS) and comparing the real observed discharge with the hydrological nowcasting scenarios. Both cases use set specific thresholds and produce a contingency table scores (based on hits, misses and false alarms).
References or other data used for validation	The validation data are the observed rainfall maps observed by the Settepani Radar located in the Liguria region for the qualitative assessment, while the Ligurian hydrometer network is used to compare the discharge in the quantitative validation process.
Skill scores	The skill score is the typical ones derived by a contingency table (as for example CSI, ROC curve, etc.)
Examples	Some studies on the PhaSt method were done (Metta et al, 2008), but the algorithm was updated and needs a new assessment section. The user interface is already available.

4.3.1 Description of the performance assessment scenario

The nowcasting rainfall field [PRD 40] is an ensemble radar-based rainfall nowcasting that predicts in the successive two hours the evolution of the two last observations of a remote sensor as a radar. The results can be visualized as probability to overcome a certain threshold of rainfall intensity or hourly accumulation.

The performance assessment will be done in two different approaches.

- Qualitatively based:

  1. *Compare the real observations of precipitation with the nowcasting maps:* The assessment will investigate if the model prediction of the areas where the precipitation is moving effectively will be hit by the event. The perfect result is when the ensemble nowcasting is similar to the observation series. In the other case, accordingly to thresholds, it will be counted the false alarms (if a portion of prediction will not be passed by the event) or the misses (if part of event not fit with the prediction);

2. *Fill a contingency table scores:* It will be reported how many times there will be presence of hit, false alarm and misses;
3. *Assessment with Skill score:* Evaluation of the typical score belonging to a contingency table as CSI, ROC curve, etc.

- Quantitatively based:

1. *Compare the real observed discharge series with the nowcasting probabilistic ensemble:* it will be verified how long the observation series lay between the probabilistic nowcasting discharge prediction envelope. The perfect result is when the ensemble prediction perfectly contains the observation series. In the other case, it will be counted the false alarms (if the real discharge is below the ensemble members) or the misses (if the real discharge is over the ensemble members);
2. *Fill a contingency table scores:* It will be reported how many times there will be presence of hit, false alarm and misses;
3. *Assessment with Skill score:* Evaluation of the typical score belonging to a contingency table as CSI, ROC curve, etc.

The assessment process can run totally automatic without an involvement of the end user. The unique partners involved is ARPAL as provider of the nowcasting products, nowcasting probabilistic discharge series and observations on the Liguria Pilot Site.

The assessment can be applied every time a new prediction of nowcasting is done (every 10 minutes) for the qualitative part, while it should be done hourly for the quantitative evaluation.

4.3.2 Description of the input data and reference datasets

In this chapter, it is described the reference observations useful to apply the assessment system as described above.

The data used in the assessment process are:

- Qualitatively based: the measurement of the Ligurian radar located 50 km far from Liguria Pilot Site;
- Quantitatively based: the real observations of the river discharge in the same sections of the prediction output. In fact, in the Liguria Pilot Site, it is available a dense hydrometer network that permits a continuous monitoring of the river flood peaks.

The data availability, according to the assessment method:

- Qualitatively based: the radar observations have a resolution of 10 minutes at the Liguria Pilot Site and are elaborated and available at ARPAL server in some minutes.
- Quantitatively based: the measures are continuous and are available at the ARPAL server (as provider of data) in real-time.

Regarding the risk in the data retrieving and countermeasures:

- Qualitatively based: the radar is maintained by the data provider to avoid any lack of information. In case of a transmission problem, the data are archived and subsequently sent to the server;
- *Quantitatively based:* The hydrometer measuring system is maintained by the data provider to avoid any lack of information. Every sensor has its system to transmit data independently and in any condition, so it is a resilient system.

In case they are not yet available, the following alternatives are applicable:

- Qualitatively based: in case of temporary unavailability of data, the comparison can be done using an interpolated map derived by the rain gauges measurements instead the radar observation. Note that in this way the comparison is done between different sensors (in term of quality and characteristics), so the results are less reliable. In case of no observations there will be no nowcasting output;
- Quantitatively based: in case of absence of the hydrometer observations, the unique possibility to assess the performance is to verify and compare possible ground effect with the observed impacts during a historical event.

4.3.3 Description of the evaluation skill scores

A contingency table is produced comparing the categories forecasts/no forecast with observations/no observations. Based on this table, the following skill scores will be computed as a function of lead time:

- Probability of Detection [POD];
- False Alarm Ratio [FAR];
- Critical Success Index [CSI].

4.3.4 Experience and examples on the Pilot Sites

Some studies on the PhaSt method were done (Metta et al, 2008), but the algorithm was updated and improved in the meantime and the evaluation of skill scores is an ongoing process. An example of the software user interface is shown in Figure 5.

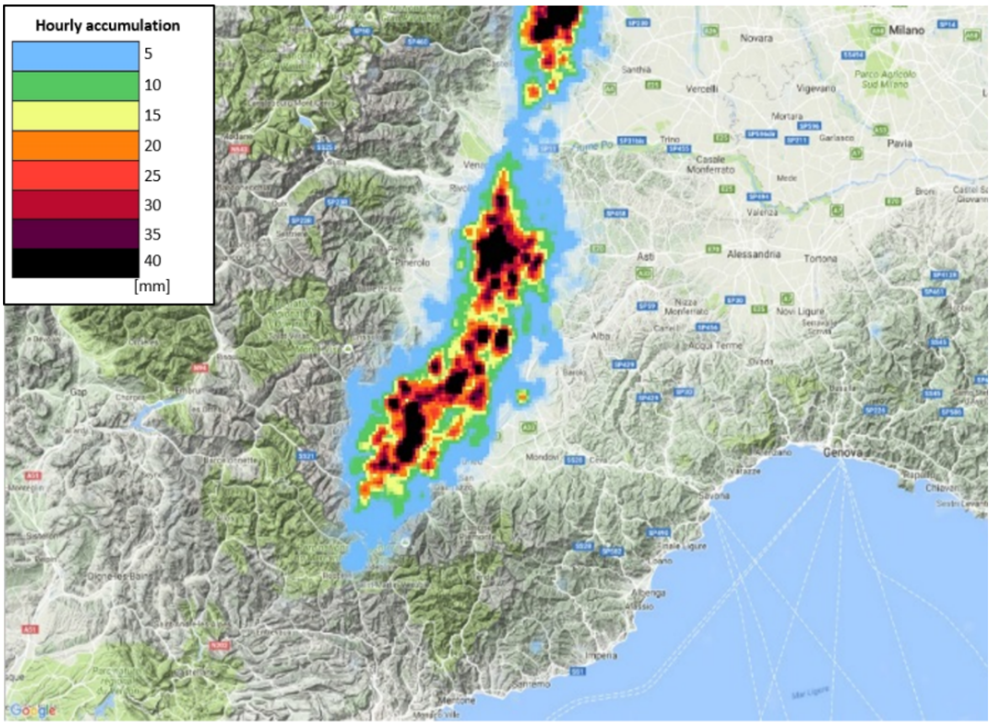


Figure 5: example of the Nowcasting rainfall field [PRD 40].

4.4 FMI - Performance assessment of Meteorological Forecasts and Nowcasts products with the pilot product A4FINN [PRD 45, 51, 170]

Table 56: Performance assessment summary for algorithm/product A4FINN.

Pilot Sites of implementation	Finland.
Description	Within the A4FINN tool, the meteorological NWP (Numerical Weather Prediction) models and LUOVA-bulletin (an official warning report of FMI) are used as direct input. The purpose of the verification is targeted to evaluate and quantify the performance of NWP model as inputs and their triggering capability with severe weather events.
Method of evaluation	The functionality of A4FINN tool is verified by assessing how the automatized risk levels induced by meteorological forecasts correspond to the actual risk level in the Pilot Site at ISTIKE.
References or other data used for validation	The reference dataset is the user assessment and the meteorological parameters provided by LUOVA bulletin for the corresponding time period (includes meteorological parameters modified by meteorology).
Skill scores	The detection skills are evaluated with three traditionally used metrics: Critical Success Index (CSI), Probability of Detection (POD) and False Alarm Rate (FAR) (Wilks, 2011).
Examples	The tool is currently active on the ISTIKE region.

4.4.1 Description of the performance assessment scenario

A4FINN is a decision-making tool, which utilizes meteorological data, an official warning report of FMI (Finnish Meteorological Institute) called LUOVA bulletin and manually-fed additional data concerning e.g. electricity outages, public events, etc. The users are duty officers of ISTIKE situation centre (four rescue services: South Savo, North Savo, South Karelia, North Karelia). Using A4FINN, the information flow from FMI will be automatically transferred to the decision makers so that with one look they will have clear picture of the situation and alert level. The A4FINN tool will give guidance for decision makers and also notify of the actions needed to be done (according the Civil Protection SOP's).

The A4FINN tool is developed from the manually fed excel-form, which has been used at ISTIKE. The defined alarm limits and thresholds for different levels of situational awareness are based on the experience of the duty officers. As in the A4FINN tool the manual part is changed to automatic version and the meteorological NWP (Numerical Weather Prediction) models are used as direct input without assessment of meteorologist, it is expected that these NWP model outputs can produce triggering of false alarms in the tool. The purpose of the verification in D3.3 is targeted to evaluate and quantify these false alarms as well as the performance of the tool to predict the coming critical events.

The functionality of A4FINN tool is verified by assessing how the automatized risk levels induced by meteorological forecasts correspond to the actual risk level in the Pilot Site at ISTIKE. It will be studied quantitatively using the traditional skill scores, e.g. FAR (False Alarm Rate) and CSI (Critical Success Index). In the specifications of the A4FINN tool, it has been requested a design of easy-using feedback feature, where the users can with minimum effort save the parameters and risk level and comment, whether these correspond each other. The verification can be performed after the A4FINN-tool is implemented to the ISTIKE site, the feedback-feature is implemented and the statistically required data amount of severe weather events is received. For summer events the minimum data, the period of summer 2018 is planned to be used; however for the winter events also winter 2018 - 2019 should be utilized.

4.4.2 Description of the input data and reference datasets

The utilized meteorological forecasts and products in A4FINN are:

- PRD 45 G Fractiles (F0, F10, F25, F50, F75, F90) of gust speeds at 10 m;
- PRD-45H Probability maps for drizzle, freezing (or supercooled) drizzle and freezing (or supercooled) rain (3 products);
- PRD 45 I The highest temperature fractile (F100) at 2 m;
- PRD 45 J The lowest temperature fractile (F0) at 2 m;
- PRD 45 K 1h precipitation (20-35 mm, 35-45 mm and >45 mm);
- PRD 45 L 24h precipitation (50-70mm, 70-120 mm and >120mm);
- PRD 45 M 1h snowfall accumulation (4-6 mm and >6 mm);
- PRD 51 Snow accumulation (24 h);
- PRD-170 Amount of hoar, dry snow, wet snow and frozen snow load on canopy.

The A4FINN tool algorithm defines the raise of the situational awareness based on the data set. The accuracy of the situational awareness level to user assessments are evaluated.

The reference dataset is the user assessment and the meteorological parameters provided by LUOVA bulletin for the corresponding time period (meteorological parameters modified by meteorology).

There are risks with this comparison. Mainly it cannot be performed before the A4FINN tool is implemented, and therefore, there is a risk that the amount of gathered incidences is statistically too small. Similar to this risk is that the feedback-feature in A4FINN is still not working and the recording of the dataset is jeopardized. The reference dataset is gathered by the end-users and it is depended on the diligence of the end-users. There is no alternative dataset existing. To reduce the risk, FMI and Finnish Ministry of Interior controls and supervises the data gathering and sets internal milestones to clarify the current situation.

4.4.3 Description of the evaluation skill scores

The detection skills at each ISTIKE region are evaluated with three traditionally metrics: Critical Success Index (CSI), Probability of Detection (POD) and False Alarm Rate (FAR) (Wilks, 2011). The contingency table with hits, false detection, misses and correct rejections can be formed based on the saved data and the reference sets. The CSI is a metric presenting the skill to detect events with correctly estimated situational awareness levels relative to all events. POD is also called the hit rate; it describes the ratio of correctly detected events to all the events and FAR is the amount of falsely estimated risk level events divided by the total amount of all the events, where risk level is not increased.

4.4.4 Experience and examples on the Pilot Sites

The image below is an example screenshot of the A4FINN tool on May 4, 2018. The tool shows the increased risk level for the ISTIKE region, because the daily precipitation (24 hours) estimate for the region is over 140 mm/24h, and this value is too high respect to expected weather. This time period would be interpreted as falsely increased risk level and the reason for the misjudgment is the product PRD 45L.

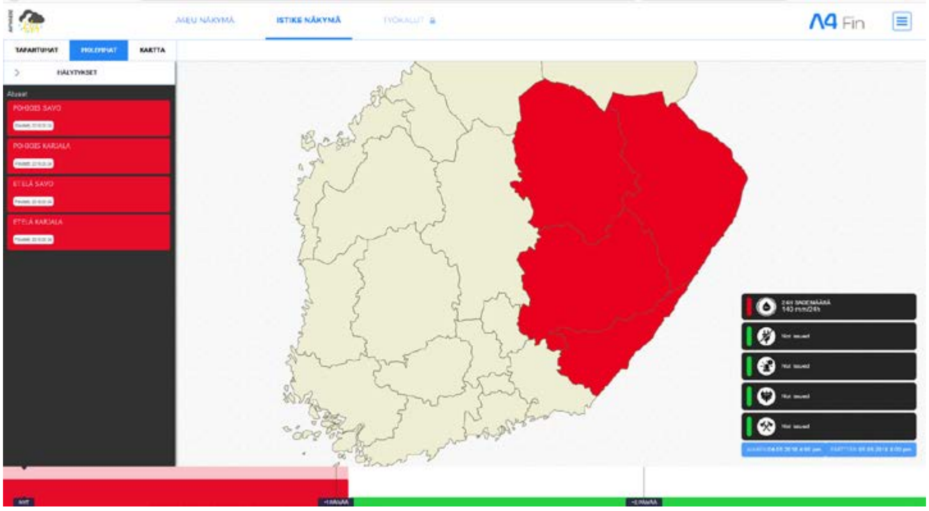


Figure 6: Example of the A4FIN automatized risk levels induced by meteorological forecasts.

4.5 ECMWF – Performance assessment of the EFAS and ERIC in real-time [PRD 80-82]

Table 57: Performance assessment summary for the EFAS and ERIC algorithm.

Pilot Sites of implementation	All Pilot Sites where forecast data are available. ERIC is not available for Finland and Norway.
Description	EFAS flood and ERIC flash-flood forecasts (real-time)
Method of evaluation	The forecasted hazard levels of EFAS and ERIC will be evaluated against a modelled hazard level (derived from the water balance for EFAS, climatology for ERIC). Where there are real-time observations of discharge, precipitation and soil moisture, or data on the observed hazard level, these will be used in the more detailed assessment of EFAS and ERIC.
References or other data used for validation	The modelled climatology of EFAS and ERIC will be used for evaluation. Observed discharge, precipitation and soil moisture for the Pilot Sites will be used for a deeper assessment of the performance if available. Reports on flash flood from the media and civil protection agencies will be used for the validation of ERIC.
Skill scores	Contingency table (hits, misses and false alarms) of the forecasted hazard levels. From the contingency table, we will calculate POD, ROC scores and symmetric extremal dependency index (SEDI).
Examples	The proposed evaluation metrics are particularly useful when evaluating extreme events (Cloke et al, 2017)

4.5.1 Description of the performance assessment scenario

EFAS flood forecasts produce runoff on a 5x5 km grid across Europe, which is then routed through a river network to produce forecasts of discharge. These discharges are compared against the modelled climatology to produce probabilities that exceed the 2, 5 and 20-year return periods. The ERIC flash flood forecasts are on a 1x1 km resolution and are a compared against the forecasted climatology to produce flash flood hazard warnings.

The evaluation of EFAS and ERIC will be done quantitatively using the above-mentioned skill scores. The evaluation will be done against the modelled climatology, which is the current standard performance measure in EFAS. We will relax the constraints in timing (+/- 1 day for EFAS and +/- 12h for ERIC) to account for model errors in the timing. We will make a qualitative assessment of spatial errors, for example if the forecast or observed event occurred in a neighbouring catchment. If there are observations of discharge, these will be added to the evaluation of EFAS, where also the model bias will be assessed. If there are observations of precipitation and soil moisture, these will be used in the evaluation of ERIC. Reports on floods and flash floods will be used in a qualitative evaluation.

If the observation data comes with a long enough time-series (at least 20 years), this can be used to derive thresholds that are comparable against the EFAS and ERIC thresholds and a full evaluation of the return period forecasts can be made. If the historical data is not available, the forecasted discharge can still be evaluated in terms of percentage of bias and skill, using for example MAE and Kling-Gupta Efficiency metrics.

The validation will be done for the operational demonstration period (October 2018 - September 2019).

4.5.2 Description of the input data and reference datasets

Forecasted flood hazards will be generated, in terms of both deterministic and probabilistic forecasts, using the 2, 5 and 20-year return period thresholds. The data for comparison will be the water balance. If available, observed values will be used.

The reference forecast will be climatology. The reference data as well as the water balance forecast. These data are made available at the same time as the forecast verifies, so the there is no difficulty in getting the data.

4.5.3 Description of the evaluation skill scores

The results will be calculated by analysing all of the forecasted and occurred flood events and then create a contingency table counting hits, misses and false alarms. This will also be done by relaxing the temporal assessment of a hit to investigate the timing error. The relaxing will be done by allowing a timing error of forecasting events (1-7 days) and see the effect on the scores. Similarly, the spatial error structure will be examined by allowing an error with increasing radius (5-50 km) of forecasting an event and see the effect on the scores. From the contingency table, the SEDI, POD and ROC scores will be calculated. The complementary scores will be MAE and Kling-Gupta efficiency.

4.5.4 Experience and examples on the Pilot Sites

EFAS routinely calculates skill scores for the EFAS domain, as an example see Figure 7. It shows the EFAS headline score, the Continuous Ranked Probability Skill Score (CRPSS) for a lead time of 10 days for the February-March period across the EFAS domain for catchments larger than 2000 km^2. The reference score is the persistence forecast. A CRPSS of 1 indicates perfect skill, 0 indicates that the performance is equal to that of the reference, and any value <0 (shown in red on the maps) indicates the skill is worse than persistence. The maps indicate that across much of Europe the forecasts are more skillful than persistence, which is also the case for other lead times. Regions shown in blue are those where EFAS forecasts are more skillful than persistence, with darker shading indicating better performance.

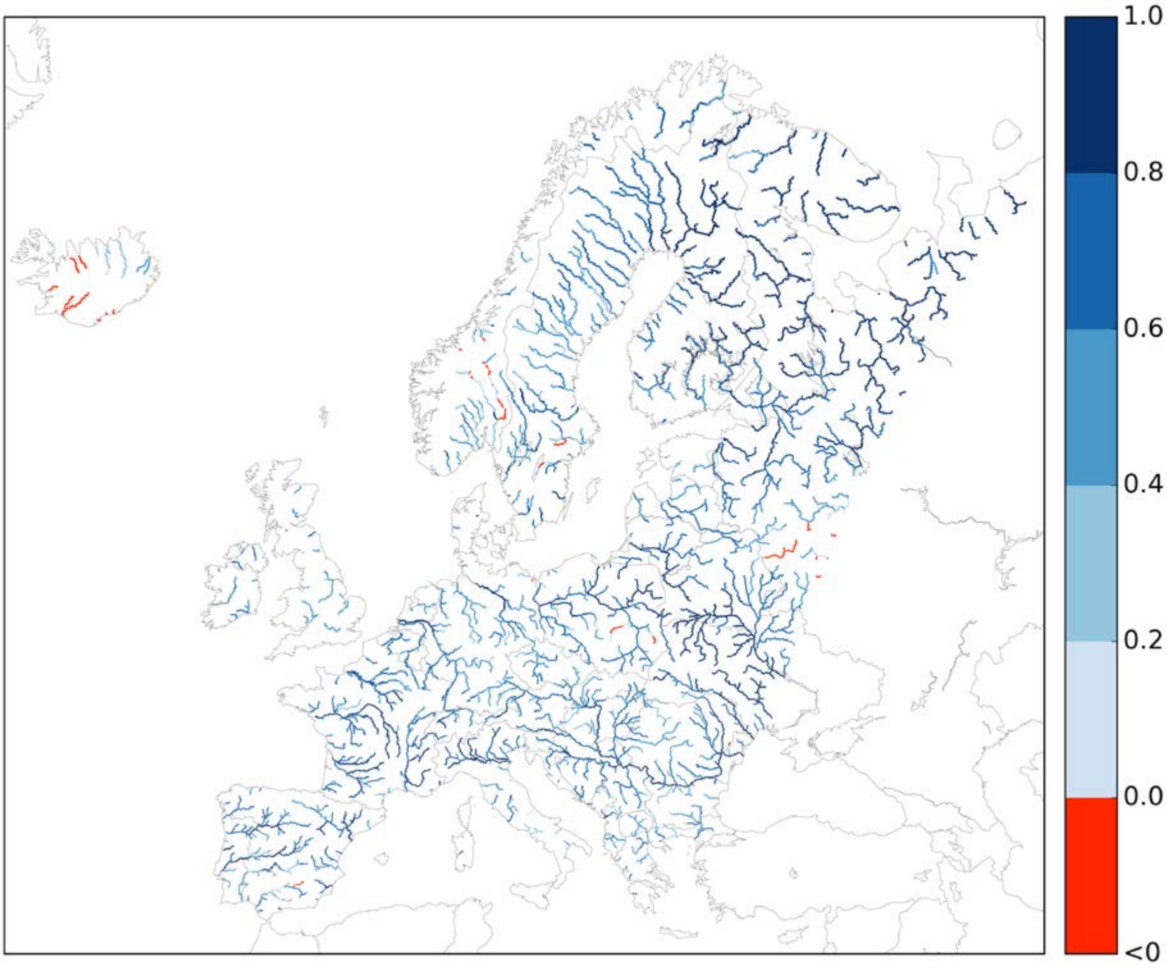


Figure 7: EFAS CRPSS at lead-time 10 days the February-March 2017 period, for catchments >2000km2. The reference score is persistence.

4.6 UPC-CRAHI - Performance assessment of the FF-EWS [PRD-93]

Table 58: Performance assessment summary for the products of the FF-EWS algorithm.

Pilot Sites of implementation	Catalonia.
Description	Estimated and forecasted flash flood (FF) hazard level.
Method of evaluation	Numerical verification of forecasts at a given observation (analysis) time and events. Qualitative validation of hazard occurrence/location
References or other data used for validation	(Deterministic) estimated flash flood hazard level at each drainage network expressed in terms of return period for lead-times up to 3 hours (at 200 m and 6-min resolution). Flash flood events reported in newspapers and floodlist.com.
Skill scores	Contingency table scores based on hits, misses and false alarms for rainfall accumulations and hazard (return period) forecasts as a function of lead time.
Examples	Numerical verification on a case occurred in a small town Agramunt (located about 100 km west of Barcelona, Spain) on 02-03 November 2015. Four died during sleep in a flooded nursing home.

4.6.1 Description of the performance assessment scenario

The FF-EWS algorithm (Flash Flood Early Warning System, Corral et al., 2009; Alfieri et al., 2017; Versini et al., 2014) implemented in ANYWHERE is designed for both monitoring and nowcasting the flash flood hazard associated with intense rainfall estimated from weather radar observations. In the Catalonia Pilot Site, the rainfall inputs are the regional composites of radar rainfall produced by the regional weather service SMC (stakeholder of ANYWHERE). Their high resolution (1 km, 6 minutes) makes them very useful to depict the evolution of localized rainfall.

For each point of the drainage network (retrieved over the Catalonia Pilot Site with a resolution of 200 m), the rainfall inputs available at a given time (both rainfall observations and nowcasts) are used to compute the basin-aggregated rainfall over a duration corresponding to the concentration time of the catchment. Using this basin-aggregated rainfall, the flash flood hazard level is determined by comparing with the values of the available Intensity-Duration-Frequency (IDF) curves assessed over the Catalonia Pilot Site for a different return period (Ciavola el al. 2017 of ANYWHERE Deliverable D2.3).

Because the quality of the input rainfall determines the FF-EWS performance, it is fundamental to validate the radar rainfall accumulations (both estimated and forecasted) with independent rainfall measurements such as those of rain gauges. In fact, the FF-EWS algorithm in the ANYWHERE MH-EWS includes a simple bias-adjustment similar to that of Park et al. (2018).

Hence, the evaluation of the FF-EWS products at the Catalonia Pilot Site will be based mainly on the comparison of forecasted flash flood return periods with those generated at a given observation (analysis) time for

- specific areas during the most significant events based on existing records and news;
- the entire domain of the Catalonia Pilot Site on a daily basis with the daily hazard summary (depicted by maximum return periods at each point of the drainage networks).

4.6.2 Description of the input data and reference datasets

The flash flood hazard (expressed in terms of return period, T) is estimated and forecasted at each point of the drainage network every 6 minutes with lead-time up to 3 hours. For a given location and time step, Figure 8a illustrates the reference return periods assessed at the time of observation (or analysis). Similarly, forecasted return periods over the drainage network are plotted for different lead-time up to 2 hours (with 30-minutes interval in Figure 8b, c, d and e).

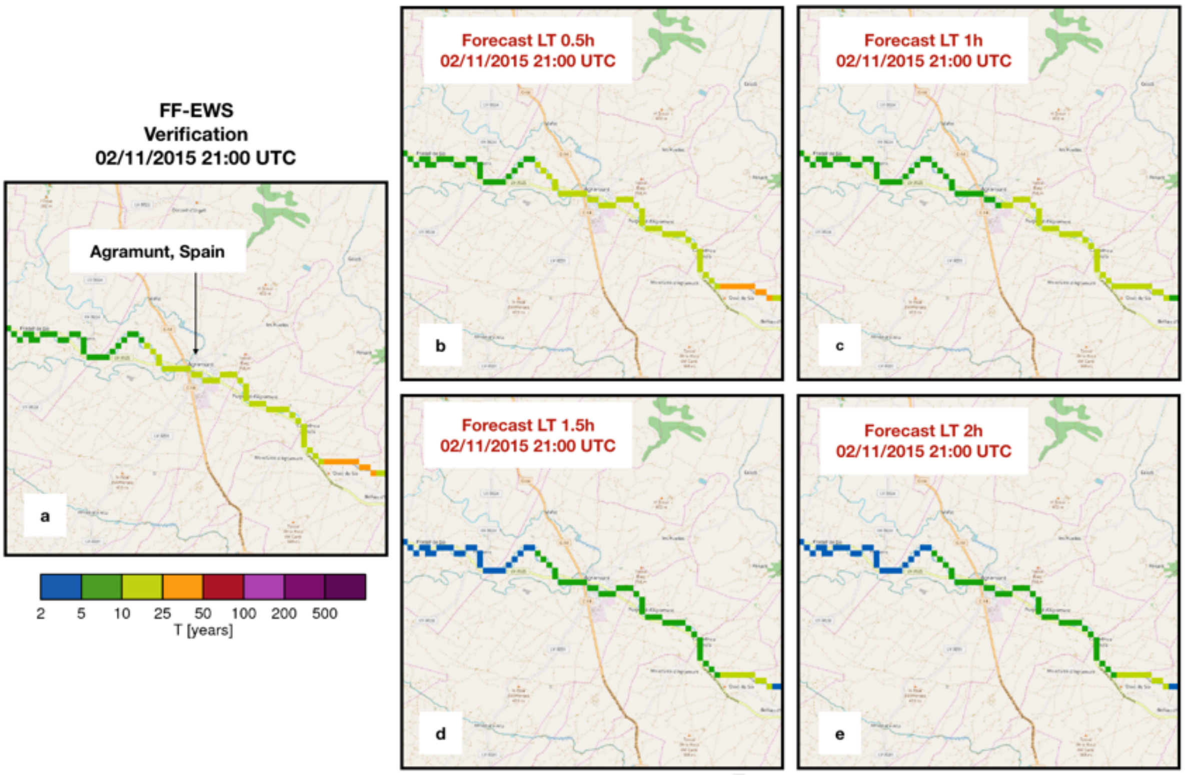


Figure 8: Example of reference data (a) and the input data (forecasted return periods as FF-hazard proxy) with forecast lead-time (LT with 30-minutes interval, b-e) around Agramunt located in the Catalonia Pilot Site.

4.6.3 Description of the evaluation skill scores

A contingency table (Table below) is created comparing the categorical "yes/no" forecasts with categorical yes/no observations. In this case, the categorical forecast is the exceedance of a given return period. Based on this table, the following skill scores will be computed as a function of lead time:

- Probability of Detection: POD=H/(H+M)
- False Alarm Ratio: FAR=F/(H+F)
- Critical Success Index (CSI; also known as threat score): CSI=H/(H+F+M)

Table 59: Standard 2x2 contingency table for dichotomous forecasts:

	Return periods Obs. (Yes)	Return periods Obs. (No)
Forecast (YES)	H (hits)	F (false alarms)
Forecast (NO)	M (misses)	Null (correct negatives)

4.6.4 Experience and examples on the Pilot Sites

According to the local news (https://youtu.be/_5uuE2wWD9g) and floodlist (http://floodlist.com/europe/spain-4-dead-river-overflows-catalonia), intense rainfalls during the day of 02 November 2015 affected the river Sió (NE Spain, also shown in Figure 8) and flooded the area around the river in the town Agramunt around late night and in the early morning of 03 November 2015. This event caused serval casualties including four deaths in a nursing home in the town center.

Given the location of the event shown in Figure 8, the FF-EWS performance can be monitored with selected time steps. Figure 9 shows a 30-min interval time series of return period (maximum extracted from a selected area of 12 km by 12 km around Agramunt) forecasts with lead-time up to 3 hours. Forecasts 30 minutes in advance are very similar to those based on observations (reference). The quality of the forecasts worsens with longer lead-times; i.e., for lead-times up to 1.5 hours, the forecasted signal can still be captured by the observation, but some large mismatches start to appear with forecasts beyond 2 hours.

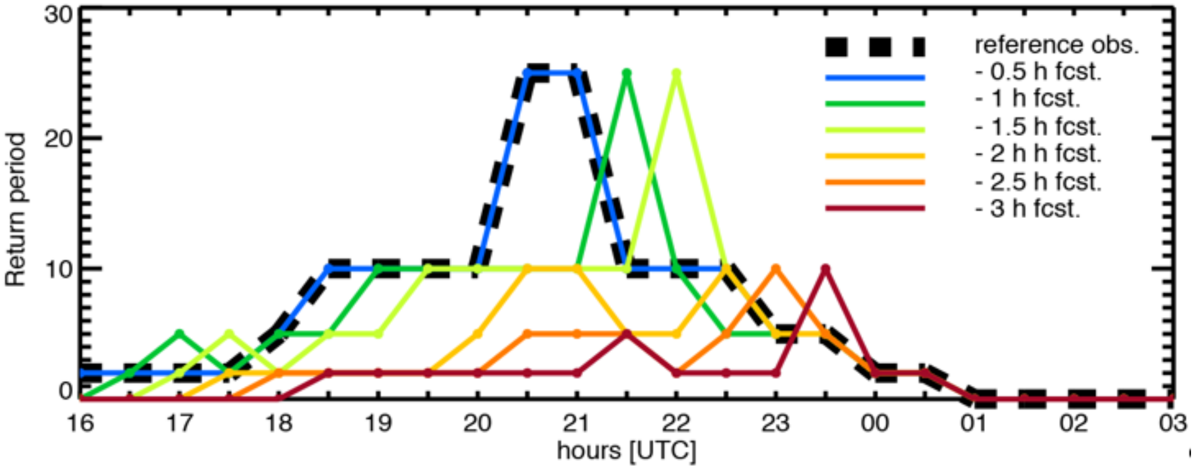


Figure 9: Time series of forecasted maximum return periods and its verification over the areas shown Figure 8 (12 km by12 km around Agramunt from 02/11/2015 16:00 to 03/11/2015 03:00 UTC).

The skill scores described in the previous section indicate the ability of the forecasts of anticipating the exceedance of certain return periods compared with the results obtained from rainfall observations. Figure 10 shows an example for the CSI (1: perfect) for the event and location presented above (Figure 9) indicating that the forecast skills beyond 1.5 hours is low.

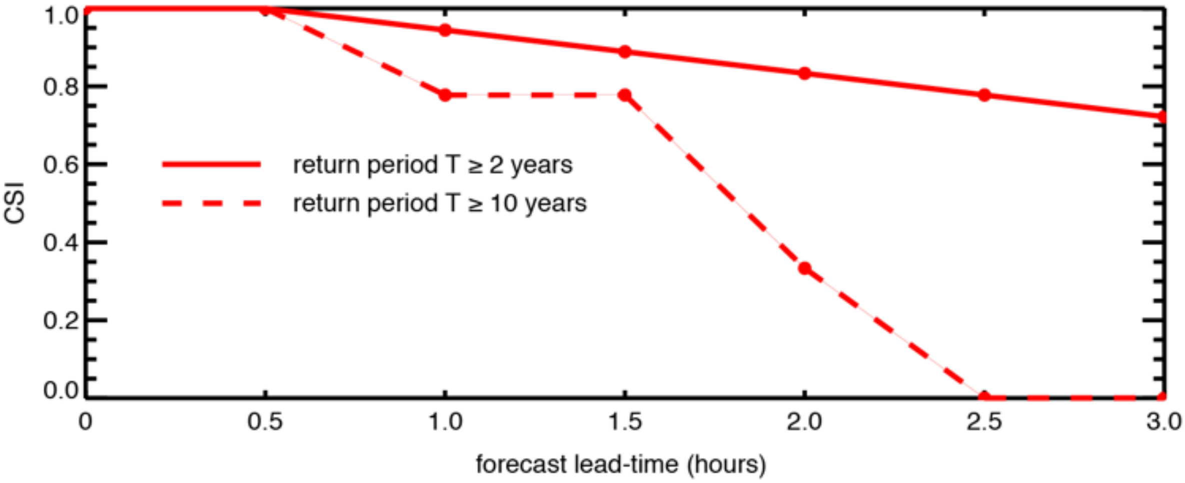


Figure 10: Critical Success Index (CSI) computed for the case presented in

Daily performance over the entire Catalonia Pilot Site will be evaluated similarly to the example shown above with the daily hazard summary to identify the areas potentially affected by flash floods and to analyse the skill at forecasting the exceedance of the 2, 10 and 100 years return periods over the entire drainage network.

4.7 CIMA - Performance assessment of the Prediction of probabilistic discharge time series on a specific section (Flood-PROOFS) [PRD 95]

As reported in the D2.3 within the Floods, flash floods, landslides and debris flows hazard, Flood PROOFS (Flood-PRObabilistic Operational Forecasting System) is one of the algorithm to be assessed.

Table 60: Performance assessment summary for algorithm Flood-PROOFS

Pilot Sites of implementation	Liguria.
Description	Flood-PROOFS (Flood-PRObabilistic Operational Forecasting System; Siccardi et al., 2005; Silvestro et al., 2011; Laiolo et al., 2013) is a system designed to assist decision makers during the operational phases of flood forecasting, nowcasting, mitigation and monitoring in small and medium catchments at regional scale (covering an area of the order of some thousands km^2).
Method of evaluation	The assessment performance is a numerical verification of how the real observed discharge in a specific hydrological section lay within the ensemble discharge prediction. Based on specific thresholds will be produced a contingency table scores (based on hits, misses and false alarms).
References or other data used for validation	The validation data are the discharge series observed by the Ligurian hydrometer network.
Skill scores	The skill scores are derived by a contingency table (as for example CSI, ROC curve, etc.)
Examples	The performance assessment will be done by measuring how long the observed discharge series lay within the forecasted probabilistic ensemble of the model. The assessment is replied every time a new run of the model is available (once a day).

4.7.1 Description of the performance assessment scenario

The Prediction of probabilistic discharge time series on a specific section [PRD 95] is the result of the modules of the Flood-PROOFS (Downscaling, Rain/snow separation, Rainfall/runoff model, etc.) that starts from the ingestion of data from different sources and managing the model workflow for hydro-meteorological forecasting.

It represents the forecast of discharge based on the meteorological input and the state variable (as Temperature, soil moisture, etc.).

The performance assessment is quantitatively based. It will be performed according these steps:

1. Compare the real observed discharge series with the probabilistic ensemble. It will be verified how long the observation series lay between the probabilistic discharge prediction envelope. The perfect result is when the ensemble prediction perfectly contains the observation series. In the other case, it will be

counted the false alarm (if the real discharge is below the ensemble members) or the misses (if the real discharge is over the ensemble members);

2. Fill a contingency table scores. It will be reported how many times there will be presence of hit, false alarm and misses;
3. Assessment with Skill score. Evaluation of the typical score belonging to a contingency table as CSI, ROC curve, etc.

The assessment process can run totally automatic without an involvement of the end user. The unique partner involved is ARPAL as provider of the probabilistic prediction of the discharge series and observations on the Liguria Pilot Site.

The assessment can be applied every time a new prediction of discharge is available (daily), using the past forecast and the observation data. Despite of the prediction horizon is 3 days the comparison will be considered only in the first 24 hours.

4.7.2 Description of the input data and reference datasets

The data used in the assessment process are the real observations of the river discharge in the same hydrological sections of the prediction output. In fact, in the Liguria Pilot Site it is available a dense hydrometer network that permits a continuous monitoring of the river flood peaks.

The measures are continuous and are available in the ARPAL server (as provider of data) in real-time.

Regarding the risk in the data retrieving, the hydrometer measuring system is maintained by the data provider to avoid any lack of information. Every sensor has its system to transmit data independently and in any condition.

In case of absence of the hydrometer observations the alternative (not optimal) possibility to assess the performance foresee to verify and compare possible ground effect with the observed impacts during historical event.

4.7.3 Description of the evaluation skill scores

A contingency table is produced comparing the categories forecasts/no forecast with observations/no observations. Based on this table, the following skill scores will be computed as a function of lead time:

- Probability of Detection [POD];
- False Alarm Ratio [FAR];
- Critical Success Index [CSI].

4.7.4 Experience and examples on the Pilot Sites

Some studies on the Flood-PROOFS method were already done (Laiolo et al, 2013), not on the Pilot Site region that will be carried out within this assessment.

An example of prediction of the probabilistic discharge time series on a specific section of the tool is shown in Figure 11.

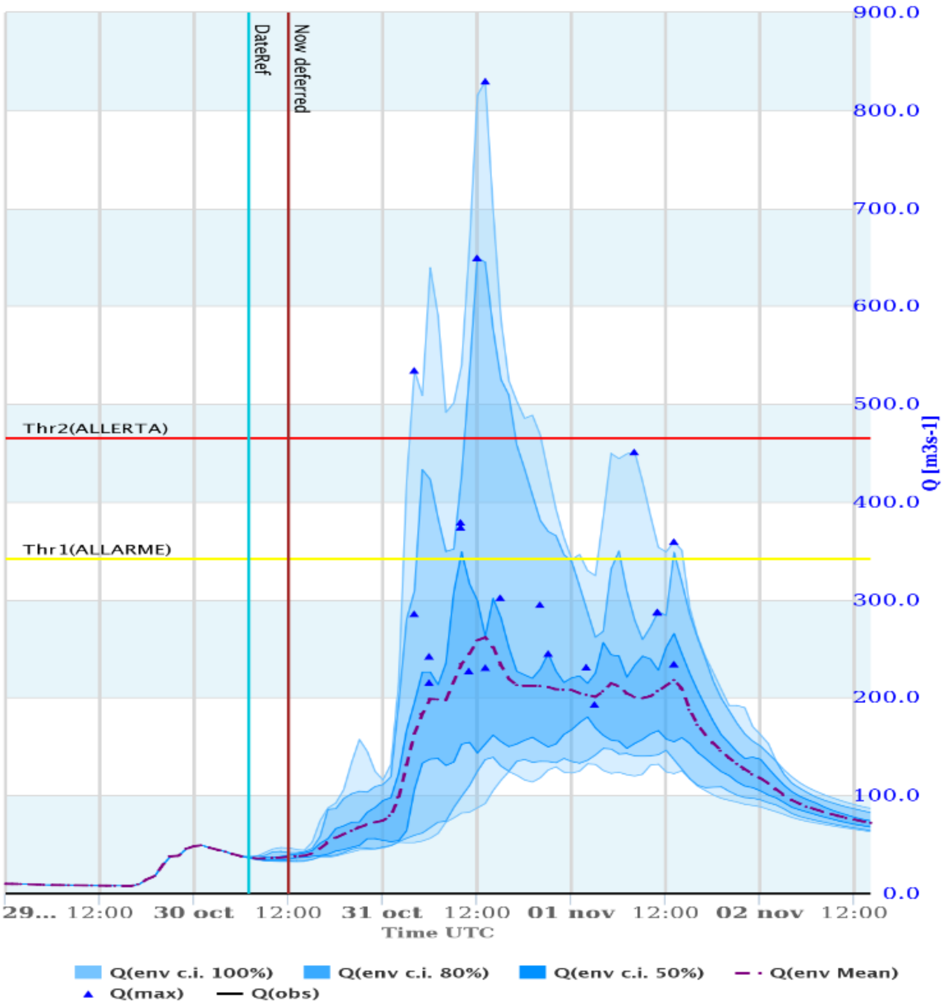


Figure 11: example of Prediction of probabilistic discharge time series on a specific section [PRD 95]

4.8 UPC-CHAHI - Performance assessment of Debris flows and Landslides [PRD-98]

Table 61: Performance assessment summary for the debris flows and landslides algorithm.

Pilot Sites of implementation	Catalonia.
Description	Estimated and forecasted debris flow and landslide hazard level.
Method of evaluation	The method of evaluation foresees: - Comparison between forecasted debris flow and landslide activity with in-situ records in monitored catchments; - Comparison between forecasted and estimated debris flow and landslide hazard level.
References or other data used for validation	Reference data used for the validation are retrieved from the following sources: - In-situ records in the Portainé and Rebaixader catchments (Hürlimann et al, 2014; Berenguer et al., 2015; Palau et al., 2017). - Debris flow and landslide events reported in the media. - Estimated debris flow and landslide hazard level at catchment scale.
Skill scores	Contingency table scores based on hits, misses and false alarms for hazard level forecasts as a function of lead time.
Examples	Comparison of the timing and magnitude of the debris flow events occurred in the Rebaixader catchment during 2010.

4.8.1 Description of the performance assessment scenario

The debris-flow and landslide forecasting algorithm (Berenguer et al., 2015) estimates the hazard level at catchment scale based on the same radar rainfall observations used by the FF-EWS (see Section 4.6).

In the Catalonia Pilot Site, the SMC radar observations are used (with resolution of 1 km and 6 minutes). Similar to the FF-EWS, the evaluation of the debris flow and landslide hazard level will be based on routinely analysis of the estimated and forecasted level throughout Catalonia, and analysis of the most significant events, especially in the monitored catchments of Rebaixader and Portainé, where the in-situ records are used to monitor the meteorological conditions and the occurrence of hyperconcentrated flows (landslides, debris floods and debris flows).

4.8.2 Description of the input data and reference datasets

The debris flow and landslide algorithm relies on the Quantitative Precipitation Estimates produced by the Meteorological Service of Catalonia, which are ingested in the MH-EWS in real time. These QPE products are used to generate precipitation nowcasts with lead-times up to 3 hours.

Part of the performance analysis compares the ability of the forecasting algorithm to anticipate the results (estimated hazard level) obtained at analysis time, based on precipitation observations.

In parallel, the results obtained in real time will be evaluated in the two monitored catchments with in-situ observations, analysing the capacity of the algorithm to identify the observed events. These observations are routinely received once a day at UPC-CRAHI Servers.

4.8.3 Description of the evaluation skill scores

The ability of the debris flows and landslides to anticipate the hazard level estimated based on observations will be estimated using the contingency table scores (POD, FAR and CSI) as a function of lead time (similarly to the FF-EWS, Section 4.6).

Besides, the frequency of significant debris flow and landslide hazard level within the catchments in the domain will be routinely reported. In the monitored catchments (and similarly for other specific events identified by reports or in the media), the analysis will focus on the ability of the system to identify them and assess their magnitudes and timing of occurrence.

4.8.4 Experience and examples on the Pilot Sites

The examples below correspond to the results obtained with the debris flows and landslides algorithm during the period between May and October 2010. As described above, the same kind of analyses will be performed during the demonstration in the Pilot Site of Catalonia (see Figure 12).

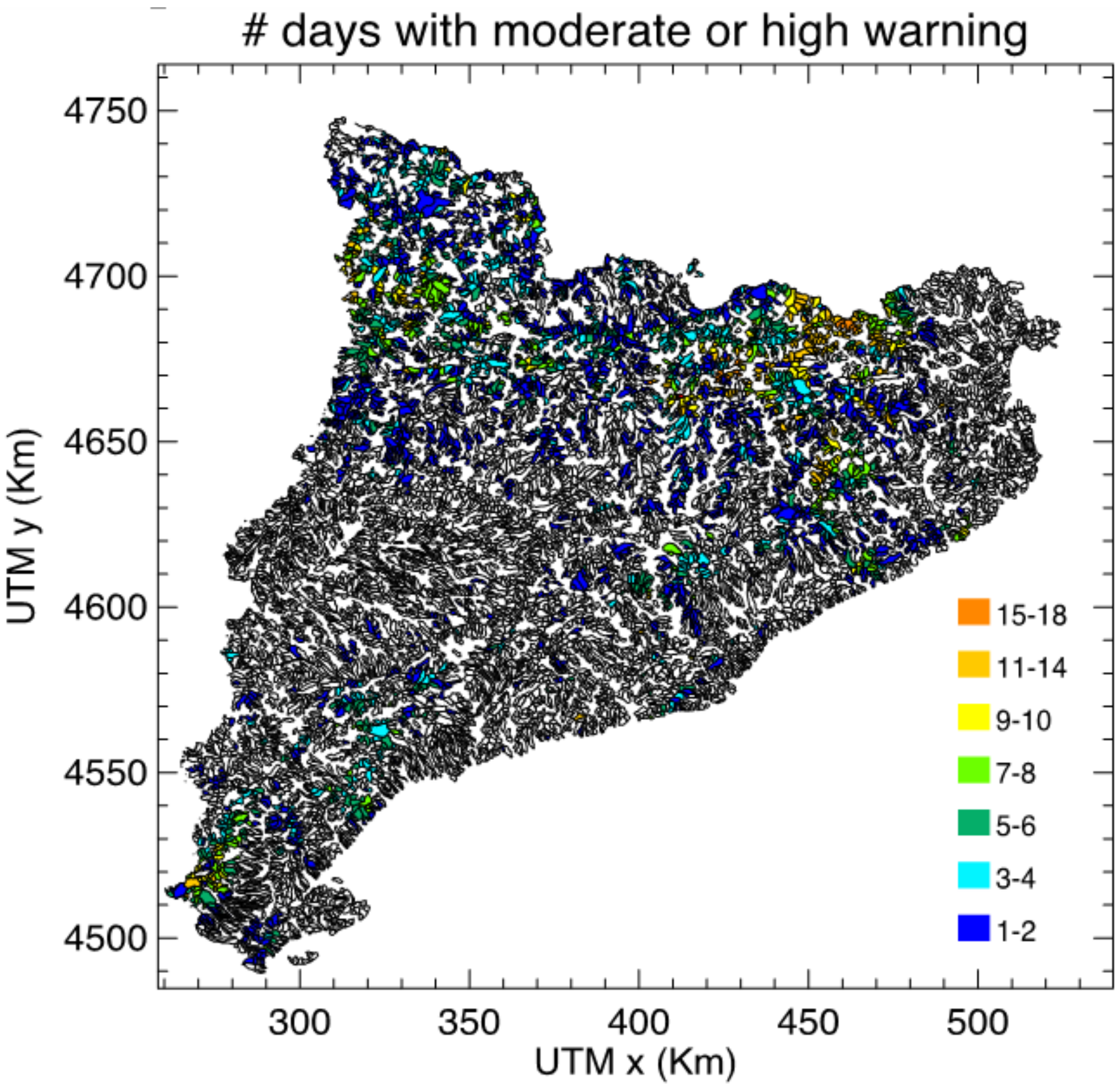


Figure 12: Number of days with significant debris flow and landslide hazard level (moderate or high) in the domain of the Pilot Site of Catalonia from May to October 2010.

Figure 13 shows the results obtained during two events in the monitored catchments of Rebaixader and Portainé. The figures show the time series of the estimated and observed rainfall in the catchments together with the time series of the diagnosed hazard level. Both cases correspond to significant events, during which the algorithm diagnosed significant hazard level (moderate or high). The analysis will focus on the timing of the signal identified by the debris flow and landslide algorithm as well as on the magnitude (i.e., how the identified hazard level matches the type of event and estimated sediment volume calculated based on in situ measurements).

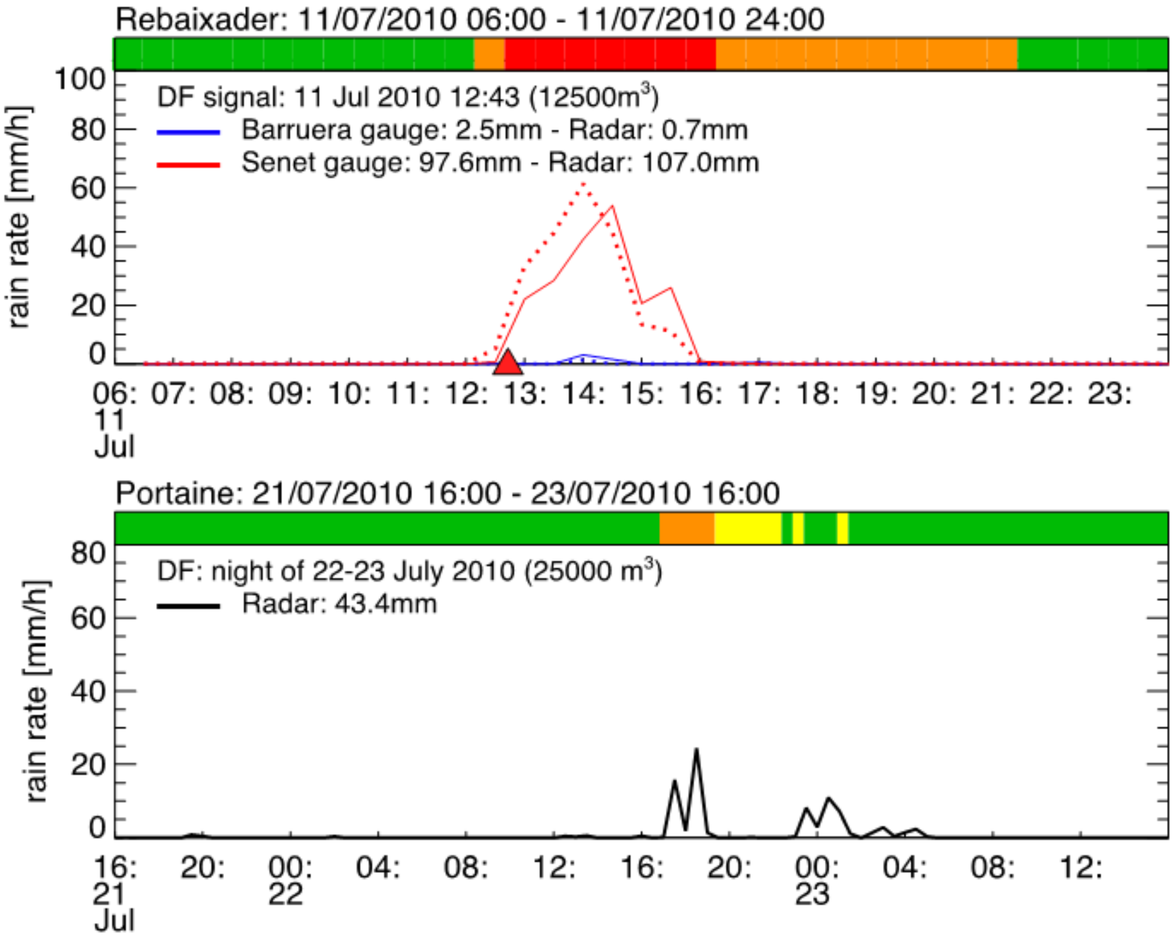


Figure 13: Time series of 30 min rain rate estimated in the Rebaixader and Portainé catchments (top and bottom, respectively). The top color bar shows the time series of the estimated hazard level obtained in the sub-basin: green, yellow, orange and red correspond, respectively, to hazard levels “very low”, “low”, “moderate” and “high”. The triangle on the x axis indicates, the beginning of debris flood event (not available for this event in Portainé) detected from geophone records. The text in the figure indicates the estimated sediment volume.

4.9 CFR - Performance assessment of the Sea Surface Level [PRD-100, 107]

Table 62: Performance assessment summary for the Regional Storm Surge model

Pilot Sites of implementation	Rogaland (Norway).
Description	Regional Storm Surge Model.
Method of evaluation	The Regional Storm Surge Model will be evaluated comparing the computed water level against water level time series available from the tidal gauge located at the Stavanger Harbour.
References or other data used for validation	Water level measurement (Norwegian Hydrographic Service).
Skill scores	Root mean squared error (RMSE), relative (%RMSE) and Pearson correlation coefficient will be used to evaluate total water level prediction.
Examples	An example of extreme event occurred in January 2017 is used to illustrate the assessment of the water level prediction of the Regional Storm Surge mode.

4.9.1 Description of the performance assessment scenario

The sea surface level is the total water (TWL) level that takes into account the astronomic and atmospheric tide. The TWL is calculated by the algorithm Regional Storm Surge Model. The evaluation of the TWL will be carried out comparing the computed against water level time series available from the tidal gauge station. The assessment of the algorithm performance to reproduce the TWL will be quantitatively evaluated. The dataset used for the model evaluation is public, therefore there is no more partner involved in the evaluation process.

The evaluation will be carried out based on past storm surge extreme event registered in the tidal gauge from 1980 onwards.

4.9.2 Description of the input data and reference datasets

The Regional Storm Surge model will run using ERA-INTERIM and High resolution forecast from ECMWF as atmospheric forcing. The data to compare the predicted result were collected by the tidal gauge station located at the Stavanger Harbour (Norwegian Hydrographic Service). They are publically available on the web site https://www.kartverket.no.

4.9.3 Description of the evaluation skill scores

The assessment of the algorithm performance to reproduce the TWL will be quantitatively evaluated in terms of the root mean square error (RMSE) relative root mean squared error (%RMSE) and Pearson correlation coefficient (*r*).

$$RMSE = \sqrt{\frac{\Sigma_i^n (\eta_O - \eta_P)^2}{n}} \quad (1)$$

$$\%RMSE = \frac{\sqrt{\frac{\Sigma_i^n(\eta_o - \eta_p)^2}{n}}}{\max(\eta_o)} \cdot 100 \quad (2)$$

$$r = \frac{\Sigma_i^n\left(\eta_o - \overline{\eta_o}\right)\Sigma_i^n(\eta_p - \overline{\eta_p})}{\sqrt{\Sigma_i^n\left(\eta_o^k - \overline{\eta_o^k}\right)^2}\sqrt{\Sigma_i^n\left(\eta_p^k - \overline{\eta_p^k}\right)^2}} \quad (3)$$

where *n* is a number of measurements in the time series at the Stavanger tidal gauge location, η_o is the observed TWL, η_p is the predicted TWL.

A value of RMSE and %RMSE closer to zero indicates a better simulation, whereas for the *r* coefficient values closer to one point to a better performance.

4.9.4 Experience and examples on the Pilot Sites

This example shows the model performance during the storm surge event occurred on 12 January 2017 in Stavanger. As can be observed in Figure 14 the model reproduces satisfactory the TWL. The skill scores calculated during this event were RMSE=0.09m; %RMSE=9.8% and *r*=0.91, showing a good performance of the model.

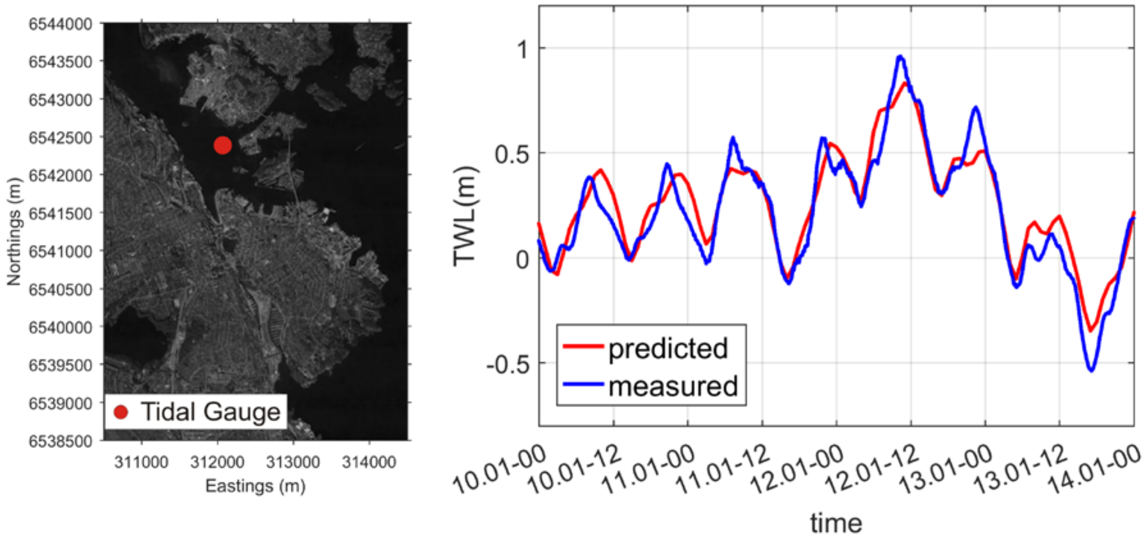


Figure 14: Regional Storm Surge model assessment: TWL predicted and measured for the extreme event occurred on 12th January 2017

4.10 CRF - Performance assessment of the wave characteristics [PRD 101-103, 108 -110]

Table 63: Performance assessment summary for the Wave Forecast

Pilot Sites of implementation	Rogaland (Norway).
Description	Wave Forecast provided by Regional Storm Surge Model.
Method of evaluation	The validation of the wave forecast provided by the Regional storm surge model will be carried out comparing the computed wave characteristic against wave time series available from satellite measurement dataset and/or comparing with different regional model outputs.
References or other data used for validation	Satellite data set (http://globwave.ifremer.fr/products/globwave-satellite-data) and Regional model outputs.
Skill scores	The skill of wave prediction will be evaluated in terms of normalized bias (NBI) and normalized root mean square error (NRMSE).

4.10.1 Description of the performance assessment scenario

The evaluation of the wave prediction of the Regional Storm Surge model is carried out comparing the computed wave against wave time series available from satellite database and/or by intercomparison with another wave model implemented in the region. The assessment of the algorithm performance for the wave forecasting will be quantitatively evaluated. The satellite data for the model evaluation are public and their area available, but the comparison with other regional models could involve to another partner during the evaluation task. The period used for the evaluation of the wave prediction expand from 2017 (dataset) onward.

4.10.2 Description of the input data and reference datasets

The Regional Storm Surge model will run using ERA-INTERIM and High resolution forecast provided by the ECMWF as atmospheric forcing, as well as water level and wave boundary conditions from the European Storm Surge Model. The references to compare the forecasted variables corresponds to the altimeter data provided by different satellites ERS-2, ENVISAT, Jason 1 and 2, Cryosat 2, and SARAL-AltiK (Queffeulou and Croizé-Fillon, 2014). They are available on the web site http://globwave.ifremer.fr. If the characteristics of the altimeter data do not fulfil the requirement for the assessment of the Regional Storm Surge model, a comparison with another regional model will be performed.

4.10.3 Description of the evaluation skill scores

The algorithm performance for the wave prediction will be quantitatively evaluated in terms of normalized bias (NBI) and normalized root mean squared error (NRMSE).

$$NBI = \frac{\Sigma_i^n (S-O)}{\Sigma_i^n (O)} (4)$$

$$NRMSE = \sqrt{\frac{\Sigma_i^n (S-O)^2}{\Sigma_i^n O^2}} (5)$$

where *n* is the number of measurements in the time series, *O* is the observed parameter (significant wave height, peak period) and *S* correspond to the simulated parameter.

NBI is an indicator of the average component of the error and a value closer to zero indicates a better simulation.

NRMSE combines information about the average and the scatter components of the error, and a value closer to zero indicates a better performance.

4.11 CRF - Performance assessment of the Local Erosion-Inundation Model [PRD-106, 111-116]

Table 64: Performance assessment summary for the Local Inundation-Erosion Model

Pilot Sites of implementation	Rogaland (Norway).
Description	Local Inundation-Erosion Model.
Method of evaluation	The Local Inundation-Erosion Model will be evaluated in terms of similarity between the prediction and extension and/or magnitude of the flooding during extreme event.
References or other data used for validation	Local information about erosion and/or extension of the inundation. The dataset depends on the measurement achieved in the Pilot Site.
Skill scores	The skill of the model will be evaluated comparing calculated area and depth of inundation with normalized root mean square error approach.

4.11.1 Description of the performance assessment scenario

The assessment of the Local-erosion model will be based on erosion and/or flooding events in Stavanger (Rogaland). The assessment of the algorithm performance will be evaluated qualitatively depending on the data availability. In order to collect the information about erosion, as well as the extension and magnitude of flooding events it will be necessary to involve to other partner during the evaluation task.

4.11.2 Description of the input data and reference datasets

The Local Erosion-Inundation model will run using water level and wave boundary conditions provided by the Regional Storm Surge Model.

The period used for the evaluation of Local Erosion-Inundation model, expand from 2017 onward, although the existence of historic information could lead to use historical extreme event.

4.11.3 Description of the evaluation skill scores

The Local Erosion-Inundation model will be evaluated by comparing the observed data about the extension and/or magnitude of the flooding during extreme event and the forecast values. The root mean square error (see eq. 1 in Par. 4.9.3) will be applied to evaluate performance of the model to reproduce the inundated area.

4.12 UoR - Air Quality and health products [PRD-117]

Table 65: Performance assessment summary for UTCI forecasts (PRD-117).

Pilot Sites of implementation	Catalonia.
Description	Medium-range forecasts of the Universal Thermal Climate Index - UTCI (PRD-117) as proxy for the assessment of the human body's comfort to heatwave-induced thermal stress.
Method of evaluation	Heatwave event defined for UTCI values above a health-meaningful threshold in terms of Heatwave timing; 2x2 contingency table, i.e. hits/missed alarms/false alarms/correct rejections.
References or other data used for validation	ECMWF reanalysis products, i.e. ERA-Interim and/or ERA5.
Skill scores	Number of hits and misses, probability of detection.
Examples	The June 2017 heatwave that hit Western Europe, especially the Iberian Peninsula.

4.12.1 Description of the performance assessment scenario

The forecast of the Universal Thermal Climate Index (UTCI) is one of the products included in the MH-EWS (PRD-117). The UTCI is a bio-climate index that uses a human heat balance model to represent the thermal stress induced by meteorological conditions to the human body (Błażejczyk, et al., 2013). In recent years, severe and prolonged episodes of summer heat such as the 2003 European heatwave have proved that extreme high temperatures are responsible for excessed mortality and morbidity in affected areas. This is due to the high heat stress levels that are commonly associated to heatwaves and have been proved harmful to human health (Di Napoli, Pappenberger, & Cloke, 2018). Forecasting UTCI gives therefore the possibility to predict the human body's comfort to heatwave-induced heat stress and potential health-related impacts.

With the MH-EWS, UTCI forecasts are provided at 6-hour time steps, with 18km spatial resolution and 10-day lead time (see Deliverable D2.3 for further details). In order to provide a meaningful heat-health warning framework, UTCI forecasts have to be reliable, i.e., to agree with observed UTCI values. The agreement between forecasts and observations is analysed via forecast verification methods (WWRP/WGNE Joint Working Group on Forecast Verification Research, 2015).

The evaluation of UTCI forecasts' performance will be done in two steps. First, the definition of a heatwave event will be explored and assessed via historical datasets, i.e., ECMWF reanalysis products will be used as a reference to determine a UTCI threshold, $UTCI_{thresh}$ hereafter, which is health-meaningful for the region. By doing so a heatwave event is treated as a *dichotomous* event, i.e., as an event that occurs when UTCI values are equal or above $UTCI_{thresh}$, and does not occur otherwise. Second, UTCI forecasts will be assessed:

- in their ability to predict the heatwave event (qualitative and quantitative evaluation);
- via a 2x2 contingency table, i.e. hits/missed alarms/false alarms/correct rejections (quantitative evaluation).

The performance assessment will focus on the summer months, i.e. 1st June to 31st August, and on the A4Cat Pilot Site. The choice of the A4Cat site is supported by the current literature which underlines the past, present and future strong relation between extreme hot summer temperatures and health impacts in the region (Tobías, et al., 2010) (Ostro, Barrera-Gómez, Ballester, Basagaña, & Sunyer, 2012).

4.12.2 Description of the input data and reference datasets

As described in the Deliverable 2.3, the UTCI forecast algorithm PRD-117 takes in input the 2m air temperature, 10m wind speed, relative humidity and radiation as predicted by the ECMWF integrated forecasting system (IFS), specifically the ensemble forecasting system. The result of the algorithm consists of UTCI forecast products in form of maps that have the same spatial resolution and lead time as IFS. The time resolution is 6 hours.

UTCI forecasts will be compared to ECMWF reanalysis products used as reference datasets. The products are: ERA-Interim (79 km x 79 km spatial resolution) and ERA5 (31 km x 31 km spatial resolution). Both ERA-Interim and ERA5 are currently released with a 3-month delay after real time. However, ERA5 is planned to be released with a 1-week delay by mid 2018 (Copernicus Climate Change Service (C3S), n.d.). ECMWF reanalysis products are the only reference that can be used against UTCI observations because their characteristics (gridded information, time steps, …) are consistent with UTCI forecasts.

4.12.3 Description of the evaluation skill scores

Forecasts of UTCI values for a given day d and time t, $UTCI_{FOREC,d,t}$, as issued on day d and the days before, $d-1$, $d-2$, … , $d-10$, will be compared with the ERA-based UTCI values at that very same day d and time t, $UTCI_{ERA,d,t}$.

As the focus is on heatwaves and heat-related stress, a $UTCI_{thresh}$ is defined from climatology (i.e., ERA products) as the UTCI reference value above which a heatwave event occurs. A heatwave is observed when $UTCI_{ERA,d,t} \geq UTCI_{thresh}$. A heatwave is forecasted when $UTCI_{FOREC,d,t} \geq UTCI_{thresh}$. With this approach, UTCI forecasts become dichotomous forecasts, i.e. yes/no forecasts. The skill of dichotomous UTCI forecasts will be evaluated with two approaches.

The first approach is based on a contingency table. A *contingency table* shows the frequency of "yes" and "no" forecasts and occurrences. The four combinations of forecasts (yes or no) and observations (yes or no), called the *joint distribution*, are:

- hit - event forecast to occur, and did occur
- miss - event forecast not to occur, but did occur
- false alarm - event forecast to occur, but did not occur
- correct negative - event forecast not to occur, and did not occur

From the contingency table the number of hits and misses will be considered to describe the UTCI forecast performance.

The second approach aims to assess, for an observed heatwave at day d, the capability of the UTCI algorithm to predict its occurrence. This is done by verifying whether $UTCI_{FOREC,d,t}$ for day d as issued by the ensemble forecasts at $d-1$, $d-2$, $\dots$, $d-10$ is also equal or above $UTCI_{thresh}$.

4.12.4 Experience and examples on the Pilot Sites

From 10th to 23rd June 2017 western and central Europe experienced the earliest heatwave of the reanalysis period (Sánchez-Benítez, García-Herrera, Barriopedro, Sousa, & Trigo, 2018). In Spain, it was the warmest June since 1965, with an average 2m air temperature 3.0 ºC above the corresponding climatological value (Agencia Estatal de Meteorología (AEMET), 2017). This exceeds by 0.1ºC the previous highest temperature value of June 2003 (Agencia Estatal de Meteorología (AEMET), 2017).

The month actually began with 2m air temperatures close to climatological values. From 7th – 8th June, however, temperatures started to increase, reaching higher-than-normal values on 10th June and peaking on the 14th and 15th June. Temperatures remained high until 18th June. On 19th June, they experienced a slight decrease, but returned to increase immediately afterwards. Temperatures decreased definitely from 25th June onwards when a low-pressure system caused below-normal temperatures in the last part of the month (Agencia Estatal de Meteorología (AEMET), 2017).

The two-week heatwave affected the whole Spain, from the south up to the northeast region of Catalonia (Agencia Estatal de Meteorología (AEMET), 2017). Figure 15 shows the heat stress levels that characterized the bioclimatic conditions of the A4Cat Pilot Site between 4th and 28th June. Moderate and very strong heat stress were prevalent, with very strong heat stress levels achieved on the heatwave peak days, i.e., 15th–16th June and 22nd–23rd June. With respect to climatology (1979-2016 reference period), the region experienced UTCI values up to 15°C higher-than-seasonal average and heat stress up to 2 categories higher-than-seasonal average.

The exceptionality of the June 2017 heatwave is also depicted by the percentiles of UTCI daily distributions over the reference period. The UTCI reached values above the 95th and 98th percentiles (about 33.6 ± 2.0°C and 34.5 ± 2.1°C, respectively[2]) close and on the heatwave peak days. Together with a minimum duration limit of 3 days,

[2] The interval is due to the variability of UTCI percentiles over June (*temporal variability*) and across the A4Cat site (*spatial variability*).

high percentiles are usually employed as thresholds to discern heatwave conditions from non-heatwave conditions. Although there is no universally accepted heatwave definition in meteorology or health-related studies, the 95th percentile is generally accepted as a defining parameter for a heatwave according to the WHO and WMO Guidance on Warning-System Development (McGregor, Bessemoulin, Ebi, & Menne, 2015). The second threshold (the 98th percentile) is employed in order to ensure that a heatwave involves at least one day with very extreme conditions; definitions of heatwaves with two thresholds are relatively frequent in bio-meteorological studies (Kyselý & Kříž, 2008). The 95th and 98th percentile thresholds defined by UTCI climatology are therefore used as $UTCI_{thresh}$ to define and identify heatwaves in the A4Cat Pilot Site. It is worth noting that percentile thresholds are grid-point specific and are thus suitable for comparison between different regions in Europe and through different time periods.

Figure 16 shows the UTCI forecasts that would have been provided by the ANYWHERE MH-EWS for the A4Cat Pilot Site if the platform was ready and operational for summer 2017. The UTCI forecasts are the mean of the ensemble UTCI forecasts as computed at 12UTC from the ECMWF IFS. A qualitative comparison between UTCI forecasts and ECMWF ERA5 reanalysis data used as proxy for observation shows the ability of the system to predict up to 3 days in advance the strong and very strong conditions of heat stress associated with the peak days of the heatwave (i.e., 14th–15th June and 21st–22nd–23rd June). The systems also captured the beginning (10th–11th June), momentary break (19th June) and end of the heatwave (25th June), where stress levels are generally lower and associated to moderate / no thermal stress conditions.

Days associated to extended, persistent, high levels of heat stress correspond to days in which extreme temperature warnings were issued for Catalonia. Specifically, "risk" levels where issued from 12th to 18th June and 20th to 25th June, and raised to "important risk" on 14th–15th, 17th June and on 22nd–23rd June (Agencia Estatal de Meteorología (AEMET), n.d.). Risk levels are based on threshold 2m air temperatures defined as the 95th percentile of the climatological temperature series for the summer months across Spain (MSSSI. Ministerio de Sanidad, Servicios Sociales e Igualdad, 2017). In Figure 17 the ability of the ANYWHERE UTCI forecast product to predict day at "important risk" for A4cat is demonstrated. The exceptional heat stress levels reached on 15th and 23rd June (i.e., UTCI values from the mean ensemble UTCI forecasts at 12UTC averaged across the A4Cat Pilot Site and above the 95th percentile) were predicted up to 6 days in advance. The signal for 14th, 17th, 22nd June becomes stronger in later forecasts. Days at "risk" were also forecast by the system using the climatological UTCI 90th percentile as threshold, as shown in Figure 18. The 90th percentile is also a reference parameter usually employed in definition of heatwave events (Perkins & Alexander, 2013). This result suggests that different climatological percentiles can be used to define and activate different warnings levels.

With respect to observation (ERA5 reanalysis data at 12UTC averaged across the A4Cat Pilot Site), the performance of the ANYWHERE UTCI forecast product during the June 2017 heatwave is assessed via the number of hits and misses. Particularly meaningful is the *probability of detection, POD* (= hits/[hits+misses]) which represents the fraction of the observed heatwave events that were correctly forecast. This measure of discrimination ranges from 0 to 1 with 1 being perfect score. Table 66 shows that when using the climatological UTCI 90th percentile as threshold, the

ANYWHERE UTCI forecast product is able to correctly predict more than half of heatwave events up to 7 days in advance (POD ≥ 56%). When using the climatological UTCI 95th percentile instead, the forecast performance is generally associated to lower POD values and shorter forecast range (up to 5 days). This might be due to the high sensitivity of POD to the climatological frequency of the event.

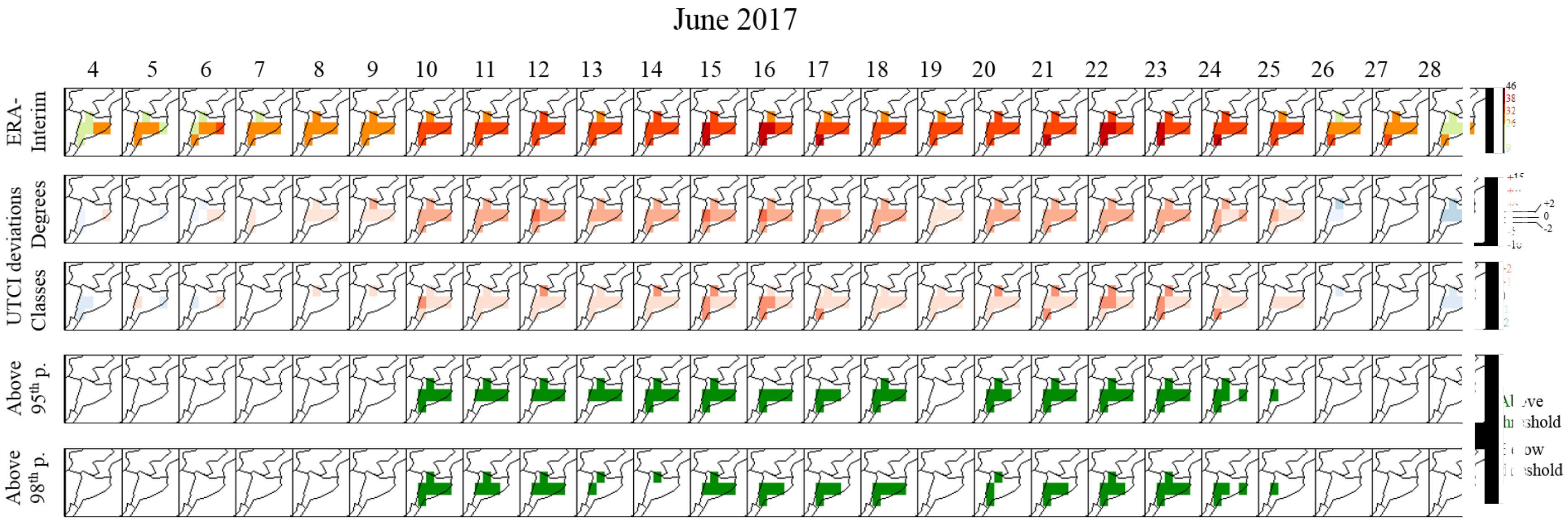


Figure 15: UTCI values at 12UTC for the A4Cat Pilot Site during the June 2017 heatwave. Upper panel: Observed UTCI values as calculated from 2m air temperature, 10m wind speed, relative humidity and radiation fields of the ERA-Interim reanalysis database, here used as proxy for observation. Second (resp. third) panel: Deviations in degrees C (resp. classes) between observed UTCI values for the indicated period and climatological UTCI values (1979-2016 reference period). Fourth (resp. fifth) panel: Grid cells with observed UTCI values above the 95th (resp. 98th) percentile of climatological UTCI values.

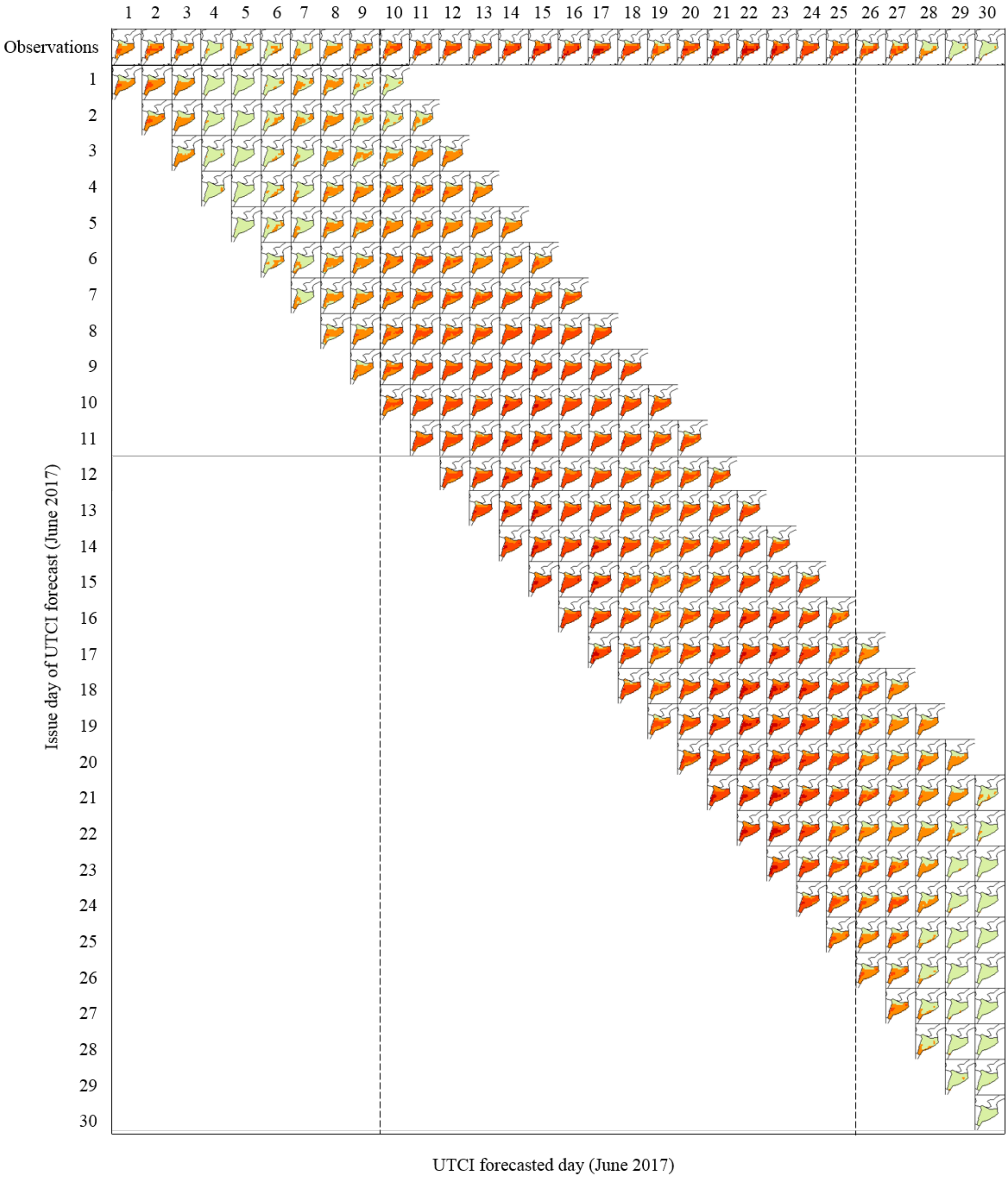


Figure 16: Ten-day UTCI forecasts (PRD-117) at 12UTC for Catalonia during the June 2017 heatwave. The top line represents UTCI values as computed at the same time point from ERA5 reanalysis database here used as proxy for observations. All other lines show forecasts issued on the days indicated on the left for the days indicated at the top.

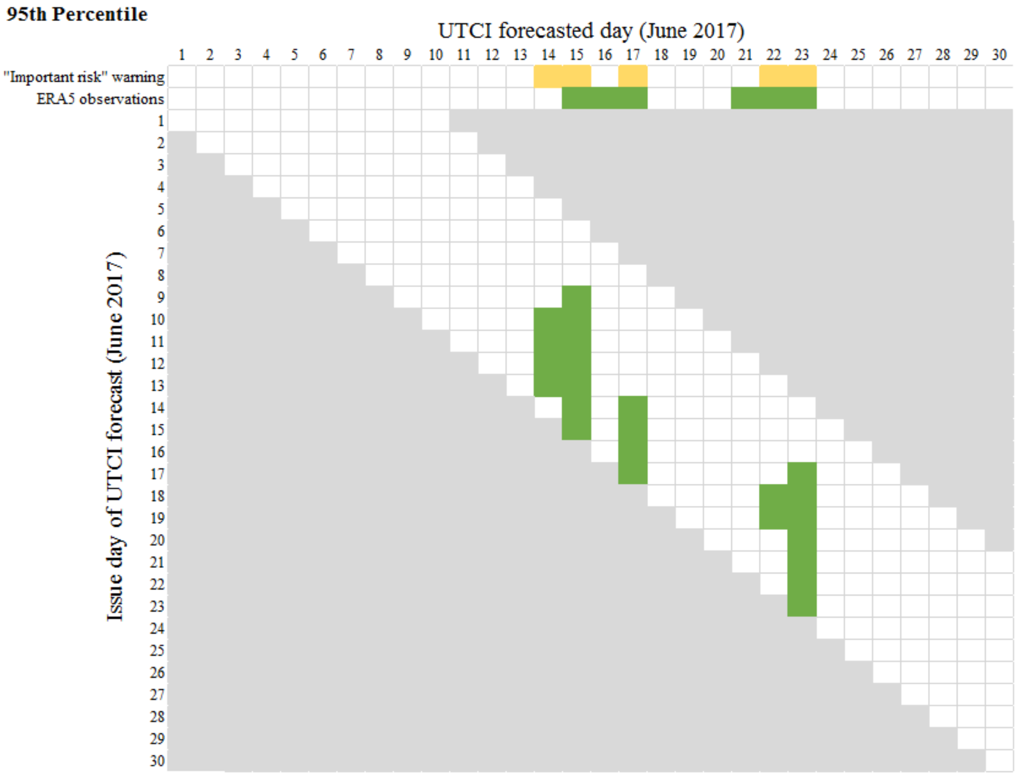


Figure 17: Predictability of the ANYWHERE UTCI forecast product (PRD-117) for Catalonia during the June 2017 heatwave. Green cells indicate days when UTCI forecasts and observations, averaged over the A4cat Pilot Site and at 12UTC, are over the corresponding 95th percentile. Yellow cells indicate days when an “important risk” warning was issued by the Spanish national weather service.

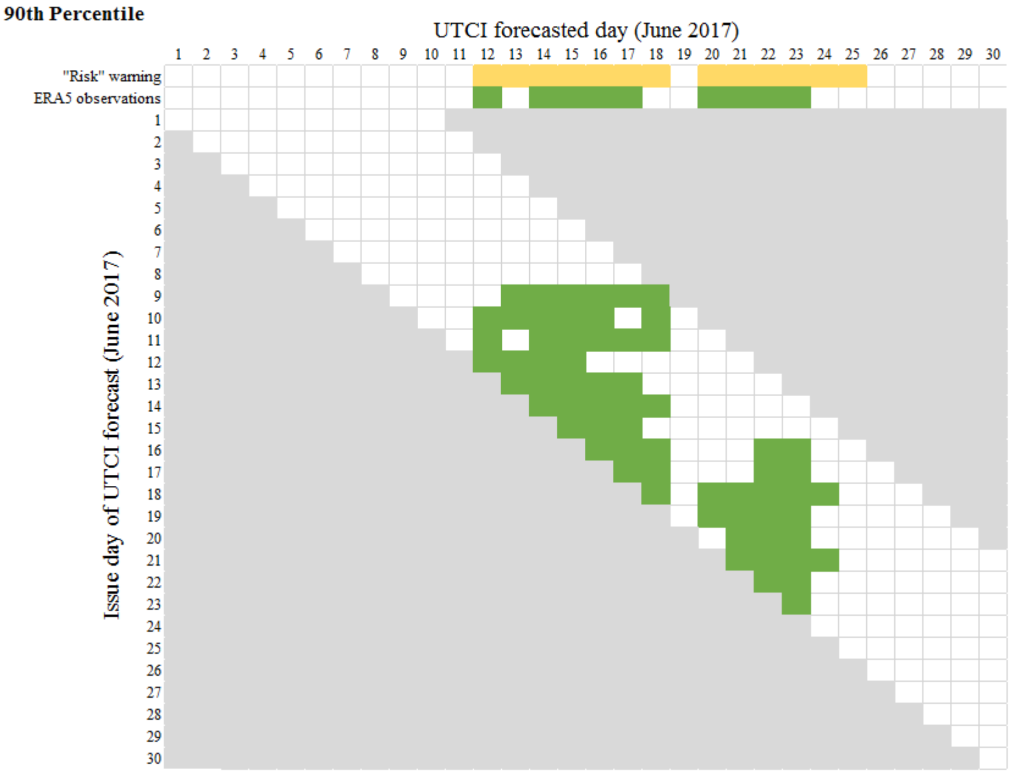


Figure 18: As Figure 17 but green cells indicating days when UTCI forecasts and observations, averaged over the A4cat Pilot Site and at 12UTC, are over the corresponding 90th percentile. Yellow cells indicate days when a “risk” warning was issued by the Spanish national weather service.

Table 66: Hits and misses of the ANYWHERE UTCI forecast product using 90th and 95th percentile levels and for different lead time (1 to 10 days). In bold forecast lead times with probability of detection POD equal or greater than 50%.

	90th Percentile			95th Percentile		
	Hits	Misses	POD [%]	Hits	Misses	POD [%]
Day 1	**8**	**1**	**89**	**3**	**3**	**50**
Day 2	**9**	**0**	**100**	**3**	**3**	**50**
Day 3	**9**	**0**	**100**	**3**	**3**	**50**
Day 4	**7**	**2**	**78**	**4**	**2**	**67**
Day 5	**5**	**4**	**56**	**3**	**3**	**50**
Day 6	**5**	**4**	**56**	2	4	33
Day 7	**6**	**3**	**67**	2	4	33
Day 8	3	6	33	0	6	0
Day 9	1	8	11	0	6	0
Day 10	0	9	0	0	6	0

4.13 ECMWF - Fire products EFFIS-GEFF algorithm [PRD-124]

Table 67: Performance assessment summary for the EFFIS-GEFF algorithm (real-time product).

Pilot Sites of implementation	Corsica, Catalonia, Liguria.
Description	Wildfire danger forecast (EFFIS-GEFF real-time).
Method of evaluation	The performance of the EFFIS-GEFF algorithm will be assessed by comparing the forecasted Fire Weather Index (FWI) to either in situ observed FWI or active wildfire hot spots captured by satellites.
References or other data used for validation	Data on Wildfire Radiative Power (observed active fires) are generated daily by the Copernicus Atmosphere Services (GFAS database), see here: http://apps.ecmwf.int/datasets/data/cams-gfas/. Temperature, precipitation, wind speed and relative humidity measured at the Pilot Sites (if available) will allow the consistency between observed and forecasted FWI to be tested.
Skill scores	A contingency table, accompanied by the estimation of probability of detection of fire event (and relative ROC curve) will be used to assess whether EFFIS-GEFF real-time provides a reliable forecast of potential fire danger in the selected Pilot Sites.
Examples	The methodology proposed here has been used to test EFFIS-GEFF reanalysis data over various countries in Europe (Vitolo et al. 2018).

4.13.1 Description of the performance assessment scenario

The European Centre for Medium-range Weather Forecasts (ECMWF) developed and now maintains the modelling engine that works as back-end of the European Forest Fire Information System: the Global ECMWF Fire Forecast (GEFF) model. EFFIS-GEFF generates two data products: a global reanalysis and a daily real-time (forecast up to 10 days ahead) dataset. Each dataset contains numerous fire danger indices, the most widely used in Europe is the Fire Weather Index (FWI), and it is used herein for the assessment of the EFFIS-GEFF algorithm. FWI is an index of potential fire danger that depends on weather conditions.

At a given location, the forecasted FWI can be assessed in two ways:

1. by comparing it with the FWI calculated from in situ weather information. This corresponds to a comparison against a common method of fire risk assessment used widely with Europe. Binning the data into categories defined by the historical percentiles of the FWI allows any expected biases in the forecasted FWI to be accounted for in the assessment.

2. If in situ measurements are not available, FWI can be compared to remotely sensed observations of active fires (also known as ‘hot spots’). In this case, particular attention is paid to the fact that high FWI values do not always correspond to active fires, but are conditional upon an ignition occurring.

The methodology to assess the performance of the EFFIS-GEFF algorithm using remotely sensed data is described in Vitolo et al. 2018. In brief, fire danger changes by country and depends on the fire climatology (the historical record of fire danger). Given a country, the thresholds that define fire danger classes are calculated from EFFIS-GEFF reanalysis. These classes are then used to re-classify the FWI forecast into 6 categories: very low, low, moderate, high, very high and extreme danger. When major fires develop and propagate for multiple days, we expect the event to be forecast-able with an FWI above the high danger threshold. For each fire detected by the Burned Areas product from the GFED4 database the corresponding FWI is extracted. If FWI is above the locally defined high danger threshold at a given cell, this is considered a 'hit' (it is a 'miss' otherwise). The number of hits and misses are summarised in a contingency table and accompanied by an estimation of probability of detection of fire events (and relative ROC curve).

However, as the GFED4 database is a composite satellite product, it is updated only few times per year which makes it unsuitable for assessing a real-time product. For this reason, EFFIS-GEFF real-time will be assessed herein using the Wildfire Radiative Power captured by satellites and made publicly available by the Copernicus Atmosphere Monitoring Service - Global Fire Assimilation System (CAMS-GFAS) on a daily basis.

It is expected that in the observation period (from June 2018 to May 2019) there will be numerous fire episodes at the Pilot Sites and, therefore, the probability of detection of fire events and relative ROC curve can be used to assess whether EFFIS-GEFF real-time provides a reliable forecast of potential fire danger in Europe.

4.13.2 Description of the input data and reference datasets

For this validation exercise, wildfire events recorded between June 2018 and May 2019 in the area of Catalonia, Liguria and Corsica will be used. The re-classified FWI forecast will be compared to the FWI calculated from locally observed weather data and/or to remotely sensed data, depending on data availability.

In case data are available at the Pilot Sites, the following variables will be needed to calculate the 'observed FWI': temperature, total precipitation, wind speed and relative humidity (all measured at the local noon). In order to bin the observed FWI into categories, the historical records of the above variables will also be needed.

The Wildfire Radiative Power, available from CAMS-GFAS (http://apps.ecmwf.int/datasets/data/cams-gfas/), will be used to assess the forecasted FWI against remotely sensed data. The FWI (reanalysis and forecast) data as well as the Wildfire Radiative Power for the regions of interest, are stored into the ECMWF internal database, therefore immediately available.

4.13.3 Description of the evaluation skill scores

Looking at events occurring in the areas of interest between June 2018 and May 2019 the following parameters will be calculated:

- The spatio-temporal extent of the fire.

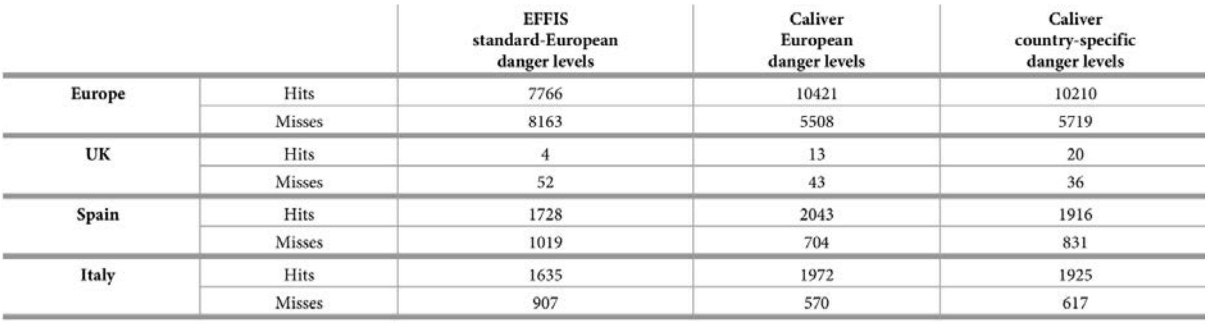
- The number of hits and misses.
- Contingency table.
- Probability of detection of fire events (and relative ROC curve).

The methodology will assess the EFFIS-GEFF realtime algorithm in a quantitative manner.

4.13.4 Experience and examples on the Pilot Sites

Numerous examples of the above methodology are available from Vitolo et al. 2018.

In Figure 19 the number of hits and misses are summarized in a contingency table and accompanied, in Figure 20, by an estimation of probability of detection of fire events (and relative ROC curve).



		EFFIS standard-European danger levels	Caliver European danger levels	Caliver country-specific danger levels
Europe	Hits	7766	10421	10210
	Misses	8163	5508	5719
UK	Hits	4	13	20
	Misses	52	43	36
Spain	Hits	1728	2043	1916
	Misses	1019	704	831
Italy	Hits	1635	1972	1925
	Misses	907	570	617

https://doi.org/10.1371/journal.pone.0189419.t012

Figure 19: Example of number of hits and misses by using different caliver thresholds

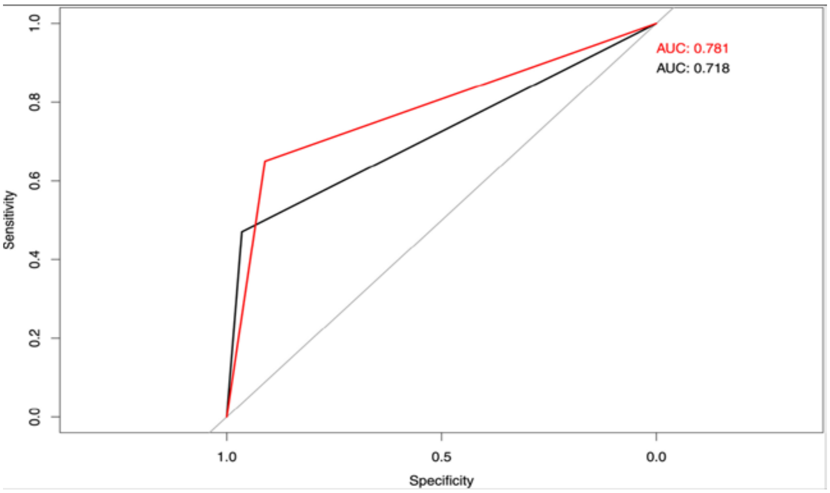


Figure 20: Example of ROC curves and AUC scores derived from the validation of EFFIS standard thresholds (black) and caliver (red) newly calibrated thresholds. https://doi.org/10.1371/journal.pone.0189419.g002

4.14 CIMA - Performance assessment of RISICO Fire Danger Rating System [PRD 140-142]

Table 68: Performance assessment summary for algorithm RISICO [PRD 140-142].

Pilot Sites of implementation	Rogaland - Norway (5km spatial resolution), Andalusia, Catalonia, Canton of Bern (250 m spatial resolution), Corsica (100 m spatial resolution), Genoa - Liguria (20 m spatial resolution).
Description	RISICO can be used both during the prevention phase and the preparedness and response phase, providing the forecast of hourly potential fire danger dynamics. Fire danger prediction can be very effective in wildland fire prevention. The capacity to identify in advance extreme weather conditions and the effect of such conditions on potential fire behaviour (rate of spread and fire line intensity), allow to put in place preventive actions able to reduce the probability of ignitions and to support decisions in preparedness phase.
Method of evaluation	The assessment performance will be based on the actual fires that will occur in Europe and respectively in all the Pilot Sites (including Andalusia) during the test phase considering a threshold on the burned area of 1000 ha at European level and locally from the local knowledge of the end users/stakeholders involved.
References or other data used for validation	The validation data are the observed burned area provided by EFFIS Rapid Damage Assessment and the information provided from the local knowledge of the end users/stakeholders.
Skill scores	The skill score includes the typical ones derived by a contingency table (as for example CSI, ROC curve, etc.) .
Examples	The performance assessment will be done measuring the number of alerts issued compared with the fire occurred considering the date of ignition, the burned area, and locally, where available, the information on prevention and firefighting activities.

4.14.1 Description of the performance assessment scenario

The Prediction of the Fire Danger is represented by the daily Fire Danger Index and by the hourly rate of spread and the effect of wind on it. It is obtained from the ingestion of meteorological data including information on topography and vegetation cover.

It represents the potential danger of a fire eventually ignited in a point of the spatial domain in a specific time in the next hours (up to 10 days in advance).

The performance assessment will be qualitatively based. It will be performed according to different steps that consist in a comparison of the prediction/observation, then a skill score evaluation based on a contingency table approach.

The assessment process can run totally automatic without an involvement of the end user at European level. At local level the involvement of the end users is essential to include information on the local prevention and firefighting activities, which can dramatically impact on the burned area.

The assessment can be applied every time a fire occurs, or prevention activities are put in place. Despite of the prediction horizon is 10 days the comparison will regard the first 72 h considering how the prediction change with respect to the time distance from the event. At the end of the fire season a comprehensive validation will be performed.

4.14.2 Description of the input data and reference datasets

The data used to run the algorithm are the meteorological forecasts provided by ECMWF, and locally for the Pilot Site of Genoa (Liguria) the meteorological forecasts provided by MOLOCH and from the local real time weather stations. The data used in the assessment process are the real observations of the burned areas and locally the prevention and firefighting activities carried out.

The EFFIS Rapid Damage Assessment are available in Near Real-Time (NRT). Locally, the information from the end users need to be provided at the end of the fire season.

The EFFIS Rapid Damage Assessment are collected automatically. Concerning local information, it depends by the capacity of the end users to collects data and provide it in time using a common protocol that needs to be defined.

In case of the lack of local information the performance assessment will be limited on the burned area provided by EFFIS Rapid Damage Assessment.

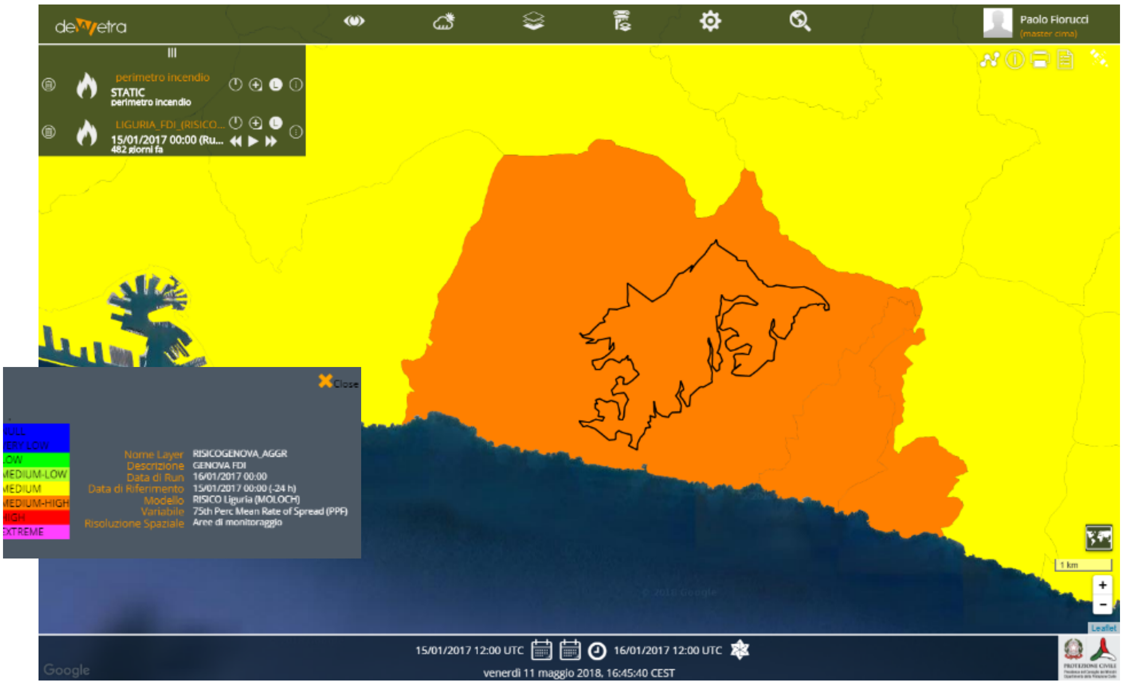
4.14.3 Description of the evaluation skill scores

The performance assessment will be quantitatively based. It will be performed according these steps:

- Compare the real burned area with the fire danger index and the rate of spread within the spatio-temporal window of the event considering also the same parameters out of the spatio-temporal window of the event;
- Fill a contingency table scores. It will be reported how many times there will be presence of hit, false alarm and misses;
- Assessment with Skill score. Evaluation of the typical score belonging to a contingency table as CSI, ROC curve, etc.

4.14.4 Experience and examples on the Pilot Sites

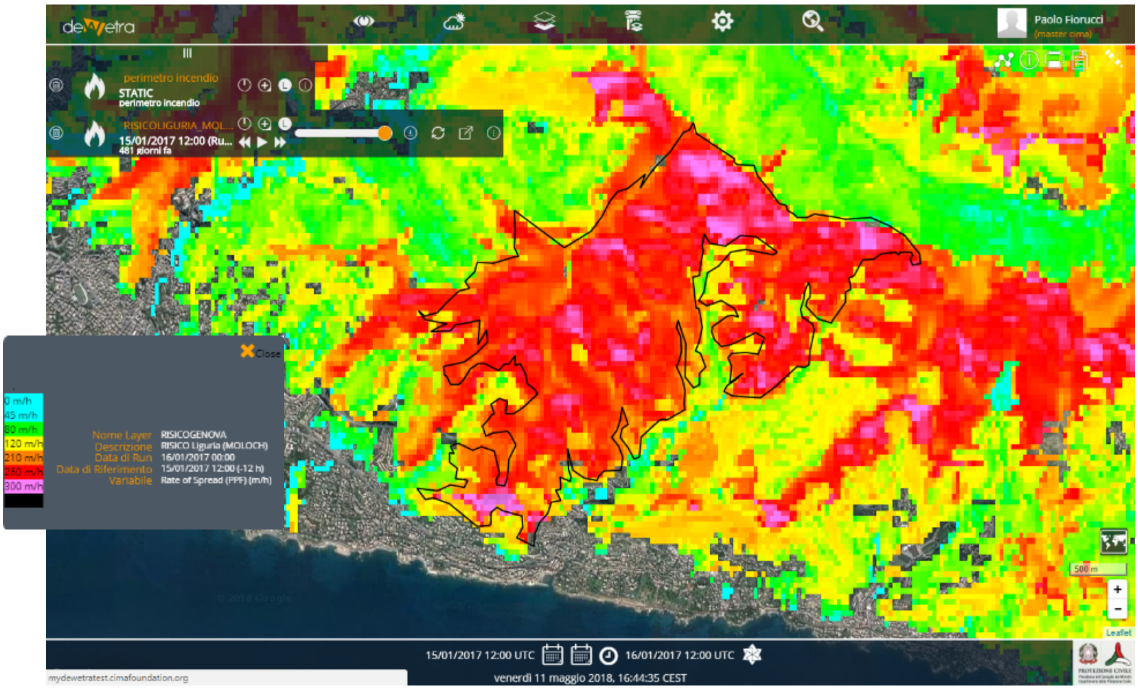
As shown in Figure 21 the fire danger index is defined in 7 classes, from low (blue) to extreme (purple). The last three levels represent respectively medium high danger (orange), high danger (red) and extreme danger (purple) and represents the different levels of alarms.



Nome Layer RISICOGENOVA_AGGR
Descrizione GENOVA FDI
Data di Run 16/01/2017 00:00
Data di Riferimento 15/01/2017 00:00 (-24 h)
Modello RISICO Liguria (MOLOCH)
Variabile 75th Perc Mean Rate of Spread (PPF)
Risoluzione Spaziale Aree di monitoraggio

Figure 21: Example of Prediction of Fire Danger Index on the Municipality of Genoa and the burned area of the fire occurred in January 2017 [PRD 138]

Besides the fire danger index provided daily and aggregated in space (sub municipality level) RISICO provide the behavior of a potential fire front in terms of rate of spread and linear intensity. In the figure below and example of the rate of spread is represented.



Nome Layer RISICOGENOVA
Descrizione RISICO Liguria (MOLOCH)
Data di Run 16/01/2017 00:00
Data di Riferimento 15/01/2017 12:00 (-12 h)
Variabile Rate of Spread (PPF) (m/h)

Figure 22: Example of Prediction of the potential rate of spread on the Municipality of Genoa and the burned area of the fire occurred in January 2017 [PRD 141]

As it is evident from the Figure 22 that the burned area is almost completely characterized by high and extreme value of the rate of spread.

The Figure 23 presents an example of ROC curves used for skills evaluation (in this case in Italy).

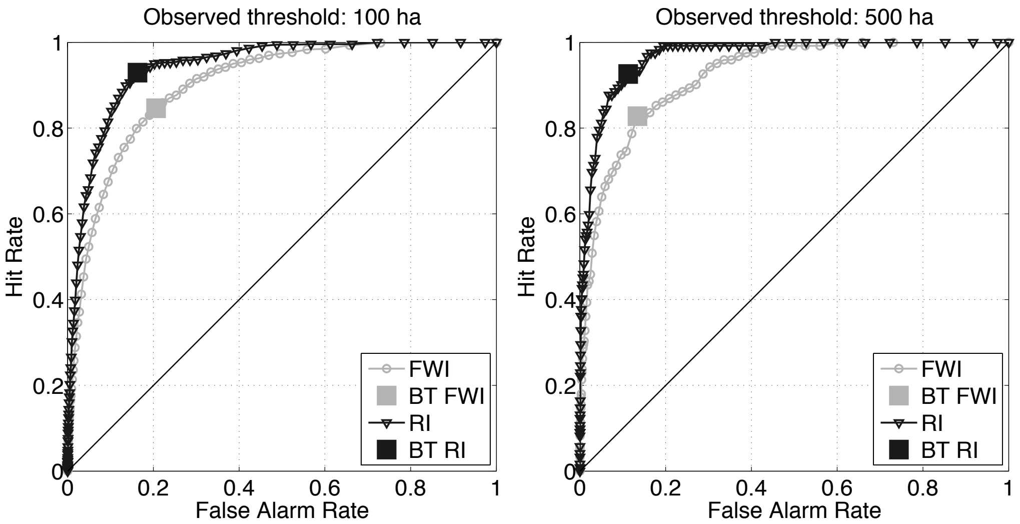


Figure 23: ROC diagram for the three forecasts for Italy and for all the available period (2007-2015) for burned area greater than 100 ha, (left) and 500 ha (right).

4.15 CIMA - Performance assessment of PROPAGATOR algorithm [PRD 143-144]

Table 69: Performance assessment summary for algorithm PROPAGATOR

Pilot Sites of implementation	Canton of Bern, Corsica, Catalonia, Liguria, Finland.
Description	The products are the dynamics of the fire perimeter and the burned area probability subject to a scenario imposed by the user (ignition point, wind velocity and direction, etc.). It represents the probability map of the burned area each 30 min of simulation. Each pixel is the number of independent simulation, which burnt the pixel itself with respect to the total number of fire simulated.
Method of evaluation	The assessment performance will be based on the actual fires simulated that will occur in the Pilot Sites.
References or other data used for validation	The validation data are the observed burned area in the Pilot Sites provided by EFFIS Rapid Damage Assessment and the information provided from the local knowledge of the end users/stakeholders.
Skill scores	The skill score is the comparison between the simulated burned area and the actual one. The performance assessment will be done measuring the difference in terms of burned area, between the simulated burned area and the actual one considering the date of ignition, and locally, where available, the information on firefighting activities. Also, the timing of the fire front will be considered.
Examples	Genoa experience in 2017.

4.15.1 Description of the performance assessment scenario

The Prediction of the burned area and the dynamics of the fire front is represented respectively by the Burned area probability map and the timing of the fire front perimeter simulated by the model. It is obtained defining the ignition point and the wind vector considering information on topography and vegetation cover.

The performance assessment will be quantitatively based. It will be performed comparing the burned area simulated and the actual one considering the timing of the fire front.

At local level the involvement of the end users is essential to run the simulation and to include information on the local firefighting activities, which can dramatically impact on the burned area. The assessment can be applied every time a fire occurs.

4.15.2 Description of the input data and reference datasets

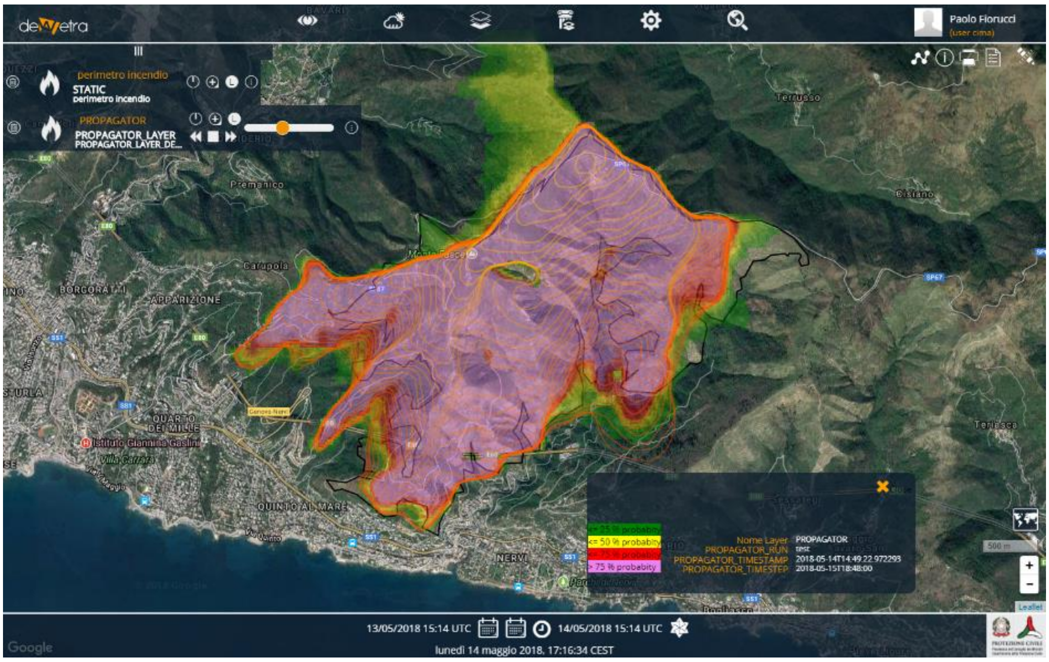
The data used by the algorithm are provided by the end user (ignition point, wind speed and direction). The data will be available during the event or just after the event. Considering that the data are available from the end users there are no risk in data retrieving.

4.15.3 Description of the evaluation skill scores

The performance assessment will be quantitatively based. It will be performed comparing the real burned area with the simulated one considering also information on the timing of the fire perimeter and, where available, firefighting activities. The difference between the actual burned area and the simulated one will provide information on the over/under estimation of the simulation considering the timing of the actual events.

4.15.4 Experience and examples on the Pilot Sites

PROPAGATOR has been tested on the event occurred in Genoa in January 2017, having available the perimeter of the burned area. In Figure 24 the prediction of the burned area is shown.



Nome Layer PROPAGATOR
PROPAGATOR_RUN test
PROPAGATOR_TIMESTAMP 2018-05-14T14:49:22.972293
PROPAGATOR_TIMESTEP 2018-05-15T18:48:00

Figure 24: Example of Prediction of the burned area of the fire occurred in January 2017 [PRD 143-144]

As it is evident from the picture the burned area is almost completely characterized by high and extreme value of the burned area probability map provided by PROPAGATOR.

Many simulations (several hundred) of PROPAGATOR have been carried out in the last months in the Pilot Sites. The ignition points of the simulations are reported in Figure 25.

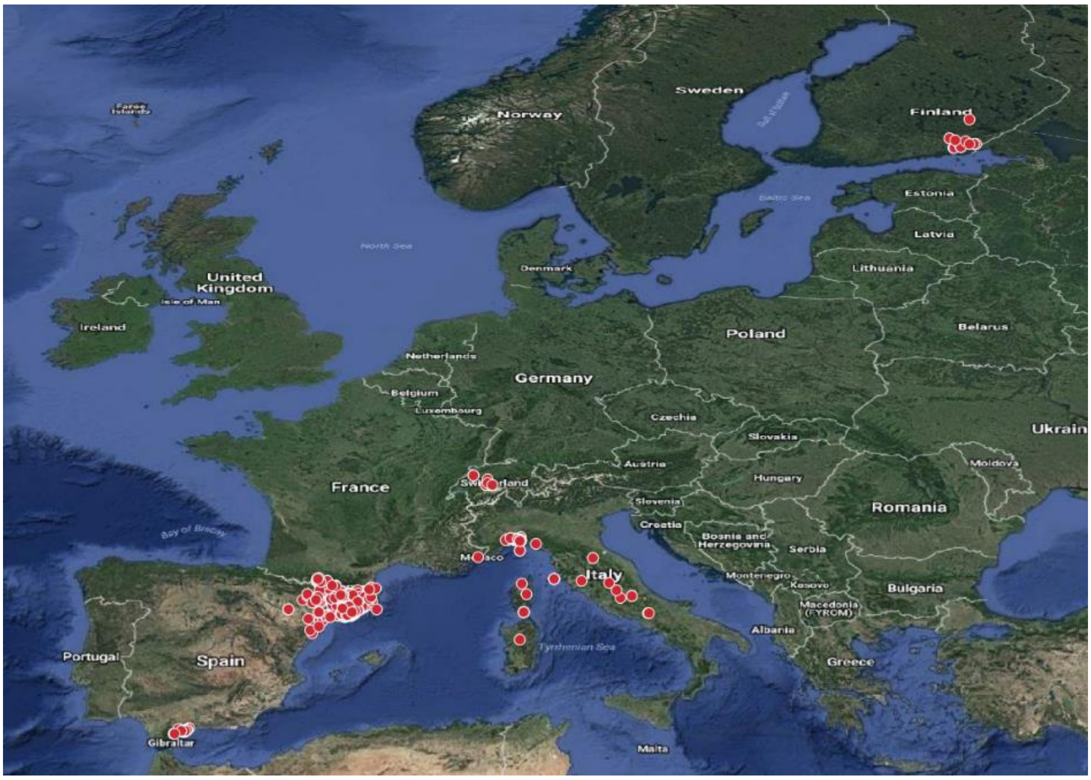


Figure 25: Ignition points of the fires simulated with PROPAGATOR in the areas where it has been implemented [PRD 143-144]

4.16 CIMA Fire Product: Performance assessment of the Fire Danger Early Warning [PRD 139]

Table 70: Performance assessment summary for algorithm RISICO feeds by real time weather observation

Pilot Sites of implementation	Liguria.
Description	It is a product of early warning fire danger. The map corresponds to the local meteorological stations that gives early warnings discretized in classes (safe, warning, alarm, no data) referred to a specific point.
Method of evaluation	The assessment performance will be based on the actual fires that will occur in the Municipality of Genoa. The number of alarms will be compared with the actual fires occurred, considering their behaviour and the prevention and firefighting activities. The assessment performance can be done, also comparing the output with direct measure on the field. For instance, the Fine Fuel Moisture Content can be compared with the measure provided by the fuel moisture sensor CS505 (Campbell Scientific) which provide measures of the moisture content.
References or other data used for validation	The validation data are the information provided from the local knowledge of the end users/stakeholders
Skill scores	The skill score includes the typical ones derived by a contingency table (as for example CSI, ROC curve, etc.). The performance assessment will be done measuring the number of alerts issued compared with the fire occurred considering the date of ignition, the burned area, and the information on prevention and firefighting activities. In addition, the quantitative assessment of the Fine Fuel Moisture Content can be done measuring the difference between the one simulated by the system and the moisture content measured by the fuel moisture sensor CS505 (Campbell Scientific) every 10 minutes.
Examples	Genoa experience in 2017.

4.16.1 Description of the performance assessment scenario

The fire danger is simulated in real time running the RISICO model feeded by real time weather observations. The output provides information on the Fine Fuel Moisture Content and the potential rate of spread, considering the effect of wind speed on the fire front. An alarm is issued when the average Fine Fuel Moisture Content in the last 12 hours is under the threshold of 7% or the rate of spread is over the threshold of 180 m/h.

The performance assessment will be qualitatively based. The performance assessment will be carried out according different steps that consist in a comparison of the simulation/observation, then a skill score evaluation based on a contingency table approach. The quantitative assessment of the Fine Fuel Moisture Content can be done measuring the difference between the one simulated by the system and the moisture content measured by the fuel moisture sensor CS505 (Campbell Scientific) every 10 minutes.

The involvement of the end users is essential to include information on the local prevention and firefighting activities, which can dramatically impact on the burned area.

The assessment should be done at the end of the operational demonstration phase for comparing the number of alerts and the actual fires occurred or during the fire season.

4.16.2 Description of the input data and reference datasets

In this chapter is described the reference observations useful to apply the assessment system as described above.

The data used to run the algorithm are the meteorological real time observations (total precipitation, wind speed, relative humidity and temperature) provided by 17 complete stations within the Municipality of Genoa.

The input and the output data are already provided in real time. The information from the end users on the occurred fires need to be provided at the end of the fire season.

In case of no fires occurred, direct measure of the Fine Fuel Moisture Content and of the fire behaviour in case of prescribed burning can be used for quantitatively performance assessment.

4.16.3 Description of the evaluation skill scores

In this chapter is better described the performance assessment process.

The performance assessment will be qualitatively based. It will be performed according these steps:

- Compare the real burned area with the early warning index and the rate of spread within the temporal window of the event considering the nearest station available;
- Fill a contingency table scores: It will be reported how many times there will be presence of hit, false alarm and misses;
- Assessment with Skill score: Evaluation of the typical score belonging to a contingency table as CSI, ROC curve, etc.

4.16.4 Experience and examples on the Pilot Sites

An analysis of the events occurred in Genoa in January 2017 has been carried out in order to identify the capability of Fire Danger Early Warning to identify the actual fire danger. Examples of the output provided by the Fire Danger Early Warning for the events occurred in Genoa in January 2017 are shown in Figure 26 and Figure 27.

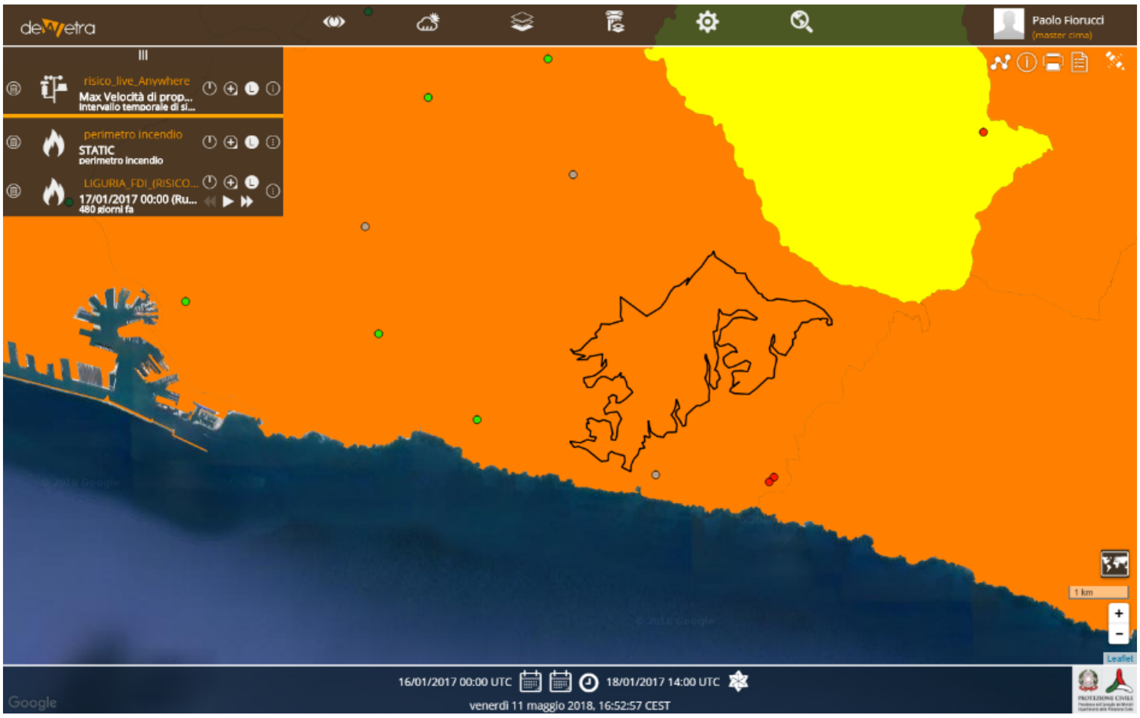


Figure 26: Example of the Fire Danger Early Warning for the events occurred in Genoa in January 2017 [PRD 139].

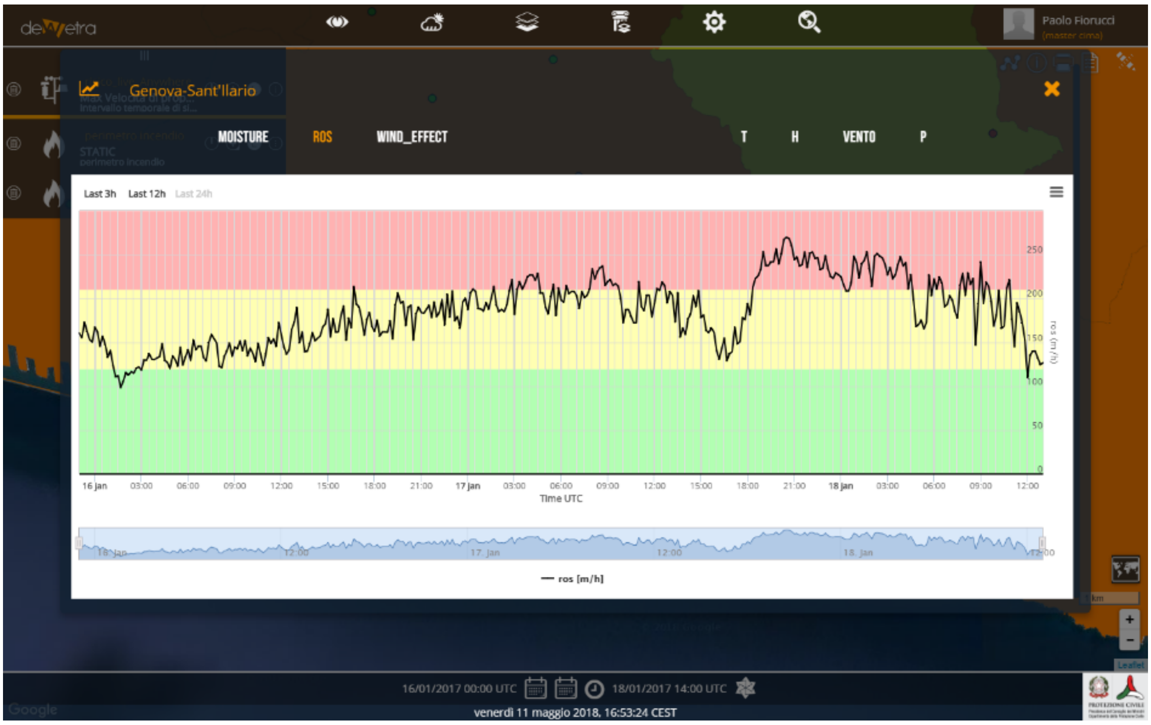


Figure 27: Example of the output provided by the Fire Danger Early Warning for the events occurred in Genoa in January 2017 [PRD 139].

4.17 WUR - Performance assessment of the Standardized Drought Indices (SPI) [PRD-148]

Table 71: Performance assessment summary for Standardized Drought Indices

Pilot Sites of implementation	Catalonia, Liguria, Corsica.
Description	The Standardized Precipitation Index (SPI) is a widely-used index to characterize meteorological drought on a range of timescales (i.e. accumulation periods). The SPI values can be interpreted as the number of standard deviations by which the forecasted anomaly deviates from the long-term observed mean. The key strength of SPI is that uses precipitation only. It characterizes drought for different accumulation periods. Long accumulation periods are a broad approximation of time availability of different water resources (e.g. soil moisture, groundwater, river discharge and reservoir storage, although standardized soil moisture, standardized groundwater, etc. are preferred).
Method of evaluation	Hind casts of SPI (each month, lead times 1-7 months) will be compared with observed monthly precipitation totals.
References or other data used for validation	In addition to recent monthly observed precipitation, long time series of precipitation data are required (at least for the period 1990-2016) to calculate the parameters of the gamma distribution.
Skill scores	Drought class based on the SPI derived from the hind casted precipitation will be compared with the class based on the SPI derived from observed precipitation.
Examples	Since September 2017, SPI is forecasted each month up to 7 months ahead for Catalonia (upstream of major reservoirs).

4.17.1 Description of the performance assessment scenario

The forecasted Standardized Precipitation Index (SPI, PRD-148) will be assessed for different accumulation periods (1, 3, 6 and 12 months). At the start of each month (e.g. June 2018) the SPI-1 will be forecasted for June, July, … December 2018 (7 months ahead) for grid cells (5 km) that cover relevant parts of the Pilot Sites. The last forecasts to be assessed likely will be from April 2019 (it is anticipated that observed precipitation May 2019 is not available before reporting). For each month in the evaluation period (June 2018 – April 2019), 7 SPI forecasts will be available, i.e. 1 month ahead up to 7 months ahead. Forecasts will be provided for precipitation accumulation periods of 1, 3, 6 and 12 months (SPI-1, SPI-3, SPI-6 and SPI-12). For instance, the SPI-1 for December 2018 only considers the forecasted precipitation in December, and the SPI-3 for December 2018 considers the forecasted precipitation in December, but also the precipitation in the previous 2 months (October, November). It depends on the accumulation period and the lead time, if the SPI-x (x>1 month) only contains forecasted precipitation. For example, the SPI-3 for December 2018 only contains forecasted precipitation (October-December) if the forecasts are done in October or earlier in 2018. However, if the SPI-3 for December 2018 is forecasted in November 2018, it will contain precipitation forecasts for November and December, but also observed precipitation in October 2018. In ANYWHERE, the observed precipitation for the forecasted SPI-3, SPI-6 and SPI-12 is obtained from the input from the LISFLOOD model (EFAS). This is available for each grid cell (5 km).

The agreement/disagreement in drought class based on the SPI derived from the forecasted precipitation (SPI forecast) and the class based on the SPI derived from observed precipitation (SPI observed) will be evaluated. Definitions of drought classes can be found in Figure 29. At the start of each month, when the monthly precipitation from the previous month becomes available, the skill can be evaluated. The forecast is perfect, if it appears that SPI forecast and the SPI observed are in the same drought class. If not, then the difference in drought class is determined. The difference is a number (skill score), hence it is a quantitative evaluation method. The maximum difference is four classes and is negative or positive indicating under-forecasting or over-forecasting.

In Catalonia, the Catalonian Water Agency (ACA) will be the partner, which will help to assess the forecasting skills. In Liguria, it will be CIMA, and in Corsica the support should be given by SIS2B. It is assumed that CIMA and SIS2B can provide the observed precipitation data (June 2018 – April 2019) and long historic time series.

4.17.2 Description of the input data and reference datasets

The input data come from the ECMWF (SEASS5) seasonal forecasts (daily precipitation), up to 7 months ahead, which are downscaled to 5 km grid for the LISFLOOD simulation (EFAS). These data are processed with the Standardized Drought Indices algorithm, which is encapsulated in the MH-EWS. The MH-EWS provides 28 SPI forecasts (four accumulation periods, seven lead times) for each month and each grid cell (5 km) in the Pilot Site. These will be compared against the observed monthly precipitation totals, which will be provided by ACA, CIMA and SIS2B. They also will provide the historic time series (from 1990 onwards) and the coordinates of the precipitation stations in the Pilot Sites.

It is worth noting that ACA, CIMA and SIS2B will make the long-term historic precipitation data available in June 2018 and that WUR will receive at the start of each month (August 2018 – May 2019) the observed monthly totals from the precipitation stations in the Pilot Sites. At the emission of this document, there is a residual risk in the data retrieving from CIMA and SIS2B (e.g. no historic data, or too short time series, data gaps) because the data requirements have not been discussed yet. There are no alternatives to adequately evaluate the skill of SPI forecasting, if the observed precipitation data are not provided.

4.17.3 Description of the evaluation skill scores

The skill scores of the forecasted Standardized Precipitation Index is derived from drought class differences between the SPI forecast and SPI observed. The skill score runs from -4 to 4 (quantitative evaluation). The median of the 51 ensemble members is taken as SPI forecast. For each grid cell in a Pilot Site, the skills will be evaluated each month, implying 28 skill scores (SPI for four different accumulation periods and

7 lead times). In this way monthly, sub-seasonal and seasonal forecasting skills can be determined.

4.17.4 Experience and examples on the Pilot Sites

Since September 2017, WUR explored together with the Catalonian Water Agency (ACA) the potential of drought forecasting for the Catalonian Pilot Site. The focus is on the forecasted drought in precipitation, through the Standardized Precipitation Index for different accumulation periods.

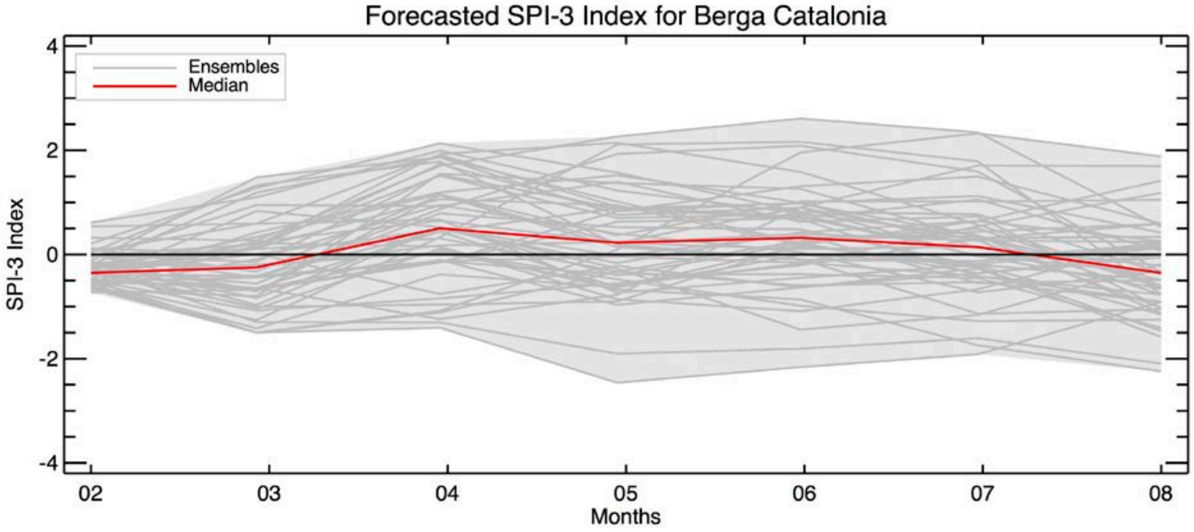


Figure 28: Forecasted probability of drought in precipitation (meteorological drought) using SPI-3 (median and 51 ensemble members) for Berga in Catalonia (7 months ahead, February – August 2018; forecast February 2018).

At the start of each month the tool forecasts the SPI for accumulation periods of 1, 3, 6 and 12 months for lead times of 1 month up to 7 months. So, for April 2018, it will be available already 7 forecasted SPI-1, SPI-3, SPI-6 and SPI-12. An example of a SPI forecast, the SPI-3 forecast done in February 2018 is given for 7 months ahead (until August 2018). The forecast includes 51 ensemble members and the median. Figure 28 provides the forecast as a time series for the particular location of Berga, which is near the Baells Reservoir. The probabilistic forecast (median) has a slightly upward trend up to April 2018, that is indicating a development towards wetter conditions. Then the SPI-3 is stable, i.e. slightly positive, which means a precipitation somewhat above the median precipitation. However, some ensemble members show that there is still a probability that a drought could develop (SPI-3 up to -2). The forecasts are also presented for the whole Pilot Site for a selected SPI (i.e. accumulation period) and lead time. Figure 29 presents, as an example, the SPI-6 forecast for April 2018 that has been done in February 2018 (lead time 2 months). The map shows the distribution of the drought classes over Catalonia.

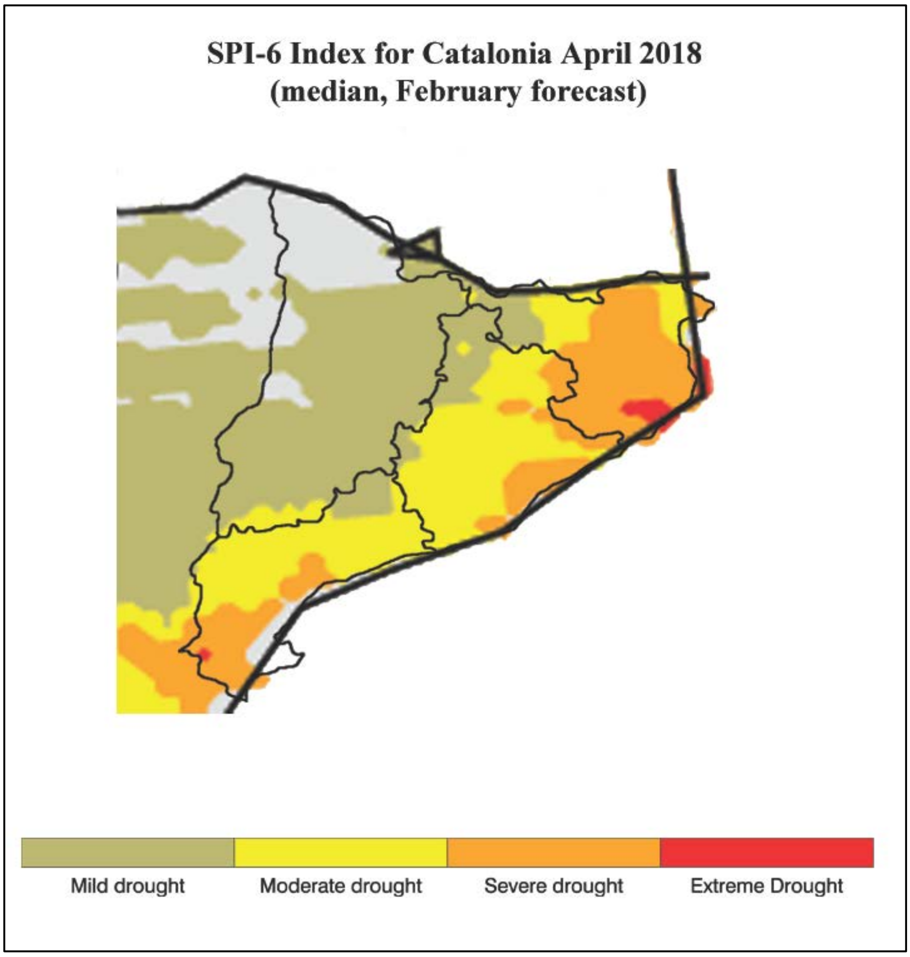

SPI-6 Index for Catalonia April 2018
(median, February forecast)
Mild drought
Moderate drought
Severe drought
Extreme Drought

Figure 29: Forecasted probability of drought in precipitation (meteorological drought) in April 2018 using SPI-6 (median of 51 ensemble members) for Catalonia (forecast February 2018). The drought classes are defined as follows: mild drought for 0>SPI≥-0.99, moderate drought for -1.00≤SPI≥-1.49, severe drought for -1.5≤SPI≥-1.99, and extreme drought for SPI≤-2.00.

The forecasted drought for April 2018 done two months ahead (Figure 29) shows that the coastal area of Catalonia has a higher probability on drought in precipitation than the western part.

In the ANYWHERE operational demonstration period, the SPI forecast will be compared against the SPI observed. This can be the time series (like Figure 28) and/or the map (Figure 29). We will inter-compare the forecasted and observed SPI for catchment areas of the following gauging stations:(i) Guardiola, (ii) Ripoll, and (iii) Masies de Roda.

4.18 WUR - Performance assessment of the Forecasted Drought in Discharge [PRD-156]

Table 72: Performance assessment summary for Threshold Drought Indices

Pilot Sites of implementation	Catalonia.
Description	Focus in the assessment is on the Forecasted Drought in Discharge (PRD-156). The Threshold Drought Method (TM) is a widely used to characterize hydrological drought, e.g. drought in groundwater or river discharge. The TM quantifies forecasted water deficits, in the case of Catalonia the deficit in discharge. A water deficit occurs when the discharge falls below the threshold. A variable threshold (VTM) will be used to account for seasonality, which means that the threshold varies over the year. The variable threshold is derived from observed discharge time series (in ANYWHERE: 1990-2016).
Method of evaluation	Hind casts of daily forecasted drought in discharge using VTM (each month, lead times 1-7 months) will be compared with water deficits derived from observed river flow in few key points relevant for water resources management under drought.
References or other data used for validation	In addition to recent observed daily discharge, long time series of observed discharge are required (at least for the period 1990-2016) to calculate the variable threshold.
Skill scores	The contingency table, including hits (H), misses (M), correct rejections (CR) and false alarms (FA). This will be applied to calculate the Brier Skill Score (BSS), if possible (sufficient data). Otherwise, the drought class will be used (see Table 71).
Examples	Monthly and seasonal skill scores for the drought in discharge: illustrated for the major drought (2006-2008) in the Ripoll River (Catalonia).

4.18.1 Description of the performance assessment scenario

The Variable Threshold Method (VTM) will be used to indicate water deficits (drought) in discharge in two key locations in the Ter and Guardiola Rivers upstream of the major dams in Catalonia, i.e. the Sau and Baells reservoirs, respectively. This method is developed based on pre-defined threshold level. The drought event starts when the discharge falls below the threshold value and ends when it rises above the threshold value. In ANYWHERE, the threshold will be derived from 80th percentile of the discharge with a centred 30 days moving average to cope with some flashiness. The drought forecasting using the VTM will be carried out as follows:

- It will be calculated the variable threshold value for each day for the discharge at the two selected locations by applying a centred moving average of 30 days using the daily time series from 1990 to 2016 (obtained from the output data from the LISFLOOD model, as a proxy for observations). This has to be done only once, rather than for every forecast.
- At the beginning of each month in the operational demonstration period (June 2018 – April 2019), it will be blended the 7-months forecasted daily discharge

for each of the 51 ensemble members data with 15 days of antecedent observed discharge data (from LISFLOOD simulation). For example, WUR will average 16-31 May 2018 daily observed discharge data with 1-15 June 2018 daily forecasted discharge data to calculate the 30-day moving average daily discharge on 1st June 2018. This means that the first day of the forecast always contains information from 15 days antecedent observations. The VTM generates for each month, location and ensemble member the drought severity (deficit in discharge) on a daily basis using the time series of 215 values. These time series will be converted into binary time series (1: drought, 0: no drought).

- Daily drought in discharge will be calculated for the 7 months using the daily binary values. This will be done at the start of every month in the demonstration period for each location and ensemble member. The daily deficits (binary) will also be computed for the observed time series (obtained from the output data from the LISFLOOD model). Hits (H), misses (M), correct rejections (CR) and false alarms (FA) will be computer for each month based on the daily values in the 7-month period.

At the start of each month, when the discharge from the previous month becomes available, the skill can be evaluated. The contingency table will be compiled, including hits (H), misses (M), correct rejections (CR) and false alarms (FA) for each month in the 7-month period, which allows to obtain skill scores for lead times from 1 to 7 months. The table will be applied to calculate the Brier Skill Score (BSS) for each month, if possible (sufficient data available). Otherwise, the drought class will be used. In that case, the forecast is perfect, if it appears that the forecasted and monthly drought in discharge are in the same drought class. If not, then the difference in drought class is determined. The difference is a number (skill score), hence it is a quantitative evaluation method. The maximum difference is four classes and is negative or positive indicating under-forecasting or over-forecasting.

For each month in the evaluation period (June 2018 – April 2019), 7 monthly skill scores will be available, i.e. 1 month ahead up to 7 months ahead. This implies that we start with the forecasts in December 2017 to obtain the 7 monthly forecasts skill scores for June 2018.

In Catalonia, the Catalonian Water Agency (ACA) will be the partner which will help to assess the forecasting skills. The skills will be assessed to relate to the water inflow to Sau and Baells reservoirs.

4.18.2 Description of the input data and reference datasets

The input data come from the ECMWF (SEAS S5) seasonal forecasts, up to 7 months ahead, which are downscaled to 5 km grid and used as input for the hydrological model LISFLOOD (revised EFAS, 2018). These data are processed with the Variable Threshold Method (VTM), which is encapsulated in the MH-EWS. The MH-EWS provides 7 forecasts of drought in discharge (seven lead times) for each month and each of the two locations in the Pilot Site. These will be compared against the drought in observed discharge, which will be provided by HYDS (obtained from the output data from the LISFLOOD model, as a proxy for observations, so-called “observations”). They also will provide the historic time series (from 1990 onwards).

HYDS in cooperation with ECMWF will make the “observations” available at the start of each month (starting with the data from December 2018 and continuing until May 2019). The risk of missing data (and/or troubles) is low, because there is already ample experience with this type of data transfer. There are no alternatives to adequately evaluate the skill of forecasting drought in discharge, if the “observed” discharge data are not provided. The actual observed river discharge data from the Ter and Guardiola Rivers cannot straightforwardly be used because of bias in the forecasted discharge.

4.18.3 Description of the evaluation skill scores

The skill scores of the drought in discharge at the two locations is derived from the Brier Skill Score (BSS), which is based upon the contingency table, including hits (H), misses (M), correct rejections (CR) and false alarms (FA), if possible (sufficient data). Otherwise, the drought class will be used (see Table 71), i.e. drought class differences between the forecasted and observed drought in discharge (similar to the SPI, Section 4.17). The skill score also runs from -4 to 4 (quantitative evaluation). The median of the 51 ensemble members is taken as forecast of drought in discharge. For each of the two locations in the Pilot Site, the skills will be evaluated each month, implying 7 skill scores (7 lead times). In this way monthly, sub-seasonal and seasonal forecasting skills can be determined.

4.18.4 Experience and examples on the Pilot Sites

The drought in discharge have been tested the seasonal skill scores for the major drought (2006-2008) in the Ripoll River (Catalonia). The Brier Skill Score is given in Figure 30.

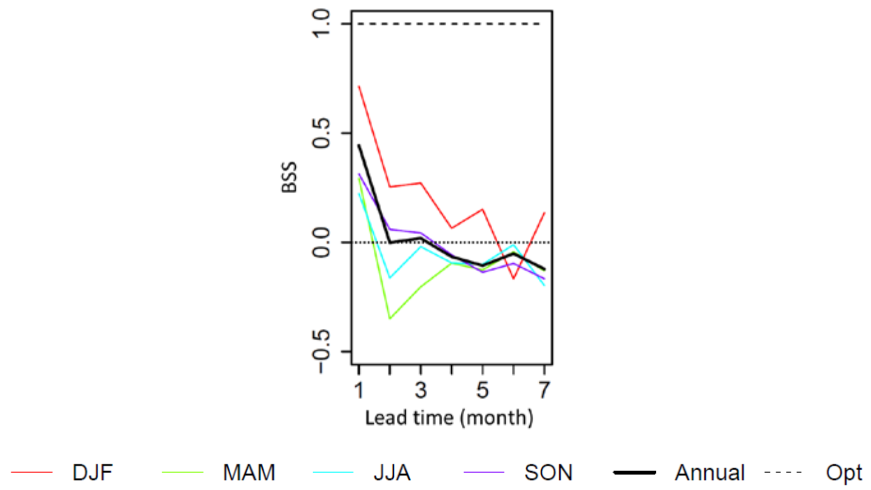


Figure 30: Forecast skill of drought in discharge of the Ripoll River (Catalonia) using the Brier Skill Score (BSS). Skills scores are determined for the major drought of 2006-2008. Skills are given for the whole year and for the different seasons. “Opt” indicates the maximum skill possible. The “0.0” line shows the forecasts based upon the climatology.

For a lead time of 1 month there is skill for the entire year using the drought forecasts provided with the ANYWHERE MH-EWS (Figure 30). For lead times of 2-3 months, skill scores are around the “0,0” line, meaning that there is only little gain of using the MH-EWS. Figure 30 shows that there are substantial differences in skills between the seasons. Skills of forecasts done in spring and summer, are limited (not beyond 1 month) relative to forecasts based upon climatology. Skills of forecasts of drought in discharge done in autumn, and particularly in winter, are considerably better. For the winter forecasts there are skills up to 5 months.

References

Agencia Estatal de Meteorología (AEMET). (2017, 09 20). *El verano de 2017, muy cálido y húmedo*. Retrieved April 2018, from http://www.aemet.es/es/noticias/2017/09/Verano_2017

Agencia Estatal de Meteorología (AEMET). (2017, July 7). *Junio, húmedo y extremadamente cálido*. Retrieved April 2018, from http://www.aemet.es/es/noticias/2017/07/Avance_Climatico_junio2017

Agencia Estatal de Meteorología (AEMET). (n.d.). *Avisos meteorológicos: Cataluña*. Retrieved from http://www.aemet.es/en/eltiempo/prediccion/avisos?w=hoy&k=cat

Alfieri, L., M. Berenguer, V. Knechtl, K. Liechti, D. Sempere-Torres, and M. Zappa, 2016: Flash Flood Forecasting Based on Rainfall Thresholds. In: Handbook of Hydrometeorological Ensemble Forecasting, Q. Duan, F. Pappenberger, J. Thielen, A. Wood, H.L. Cloke, and J. C. Schaake, Eds., Springer Berlin Heidelberg, 1-38.

Applications. Springer-Verlag New York, 222 pp. ISBN: 978-1-4614-9298-6.

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.

Biondi G., D'Andrea M., Fiorucci P., Gaetani F., Negro D., 2010: PROPAGATOR: a rapid and effective tool for active fire risk assessment, Geophysical Research Abstracts Vol. 12, EGU2010-10877-1, EGU General Assembly.

Błażejczyk, K., Jendritzky, G., Bröde, P., Fiala, D., Havenith, G., Epstein, Y., . . . Kampmann, a. B. (2013). An introduction to the Universal Thermal Climate Index. *Geogr Pol, 86*(1), 5-10.

Ciavola, P., Fernandez Montblanc, T., Armaroli, C., Berenguer, M., Bergman, T., Cloke, H., Corral, C., Di Napoli, C., Dotttori, F., Duo, E., Gascon, E., Harri, Kalas, M., A-M, Koistinen, J.,Láng, I., von Lerber, A., Llort, X., Park, S., Pylkkö, P., Rebora, N., Rodríguez Ramos, Á., Salamon, P., Sempere, D., Smith, P.J., Van Lanen, H.A.J., Sutanto, S.J., Taufik, M., Teuling, A.J., Uijlenhoet, R., Vitolo, C. and Vousdouskas, M., 2017: Report describing the improved version of the algorithms to be incorporated into the MH-EWS. ANYWHERE Internal Report, Ferrara, Italy (Confidential).

Cloke, H. L., Pappenberger, F., Smith, P. and Wetterhall, F. (2017) How do I know if I've improved my continental scale flood early warning system?, Environmental Research Letters, 12(4), https://doi.org/10.1088/1748-9326/aa625a

Copernicus Climate Change Service (C3S). (n.d.). *Climate reanalysis*. Retrieved April 2018, from https://climate.copernicus.eu/products/climate-reanalysis

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: Flood Risk Management: Research and Practice, P. Samuels, S. Huntington, W. Allsop, and J. Harrop, Eds., Taylor & Francis, 1295- 1303.

Di Napoli, C., Pappenberger, F., & Cloke, H. (2018). Assessing heat-related health risk in Europe via the Universal Thermal Climate Index (UTCI). *International Journal of Biometeorology*: https://doi.org/10.1007/s00484-018-1518-2

Fiorucci P., D'Andrea M., Negro D., Severino M. (a cura di), 2011, Manuale d'uso del sistema previsionale della pericolosità potenziale degli incendi boschivi RIS.I.CO., Presidenza del Consiglio dei Ministri – Dipartimento della protezione civile e Fondazione CIMA

Fiorucci P., Gaetani F., Minciardi R., (2008). Development and application of a system for dynamic wildfire risk assessment in Italy. Environmental Modelling & Software.

Elmore, K., Grams, H., Apps, D., and Reeves, H. (2015): Verifying forecast precipitation type with mPING. Wea. Forecasting, 30, 656–667, https://doi.org/10.1175/WAF-D-14-00068.1.

Fehlmann, M., Gascón, E., Rohrer, M., Schwarb, M., Stoffel, M. (2018). Estimating the snowfall limit in alpine and pre-alpine valleys: A local evaluation of operational approaches, Atmos. Res., 204, 136-148, https://doi.org/10.1016/j.atmosres.2018.01.016.

Gascón, E., Hewson, T., Haiden, T. (2018). Improving predictions of precipitation type at the surface: description and verification of two new products from the ECMWF ensemble. *Weather and Forecast*ing, 33, 89–108.

Gascón, E., Hewson, T., Sahin, C. (2018). New products for precipitation type probabilities, ECMWF Newsletter. No. 154, 2-3.

Germann, U., & Zawadzki, I. (2002). *Scale-dependence of the predictability of precipitation from continental radar images. Part I: Description of methodology.* Monthly Weather Review.

Guillot, P., and D. Duband, 1967: La methode du Gradex pour le calcul de la probabilite des crues a partir les pluies. AISH Publ. 84, 560–569.

Hürlimann, M., C. Abancó, J. Moya and I. Vilajosana, 2014: Results and experiences gathered at the Rebaixader debris-flow monitoring site, Central Pyrenees, Spain. Landslides, 11, 939-953.

INSPIRE Cross Thematic Working Group on Observations & Measurements. (n.d.).

Kyselý, J., & Kříž, B. (2008). Decreased impacts of the 2003 heat waves on mortality in the Czech Republic: an improved response? *International Journal of Biometeorology, 8*, 733–745.

Laiolo, L., S. Gabellani, N. Rebora, R. Rudari, L. Ferraris, S. Ratto, H. Stevenin, and M. Cauduro, 2013: Validation of the Flood-PROOFS probabilistic forecasting system. Hydrological Processes. http://dx.doi.org/10.1002/hyp.9888, 2013

McGregor, G., Bessemoulin, P., Ebi, K., & Menne, B. (2015). *Heatwaves and Health: Guidance on Warning-System Development.* Geneva, Switzerland: World Meteorological Organization.

Metta, S., von Hardenberg, J., Ferraris, L., Rebora, N., & Provenzale, A. 2009. Precipitation nowcasting by a spectral-based nonlinear stochastic model. Journal of Hydrometeorology, 10(5), 1285-1297.

MSSSI. Ministerio de Sanidad, Servicios Sociales e Igualdad. (2017). *Plan Nacional de actuaciones Preventivas de los efectos de los excesos de temperaturas sobre la salud.* Retrieved from https://www.msssi.gob.es/ciudadanos/saludAmbLaboral/planAltasTemp/2017/docs/Plan_Nacional_de_Exceso_de_Temperaturas_2017.pdf

Open Geospatial Consortium. (2014). http://www.opengeospatial.org/. Retrieved 2016, from opengeospatial: https://portal.opengeospatial.org/files/?artifact_id=52803

Ostro, B., Barrera-Gómez, J., Ballester, J., Basagaña, X., & Sunyer, J. (2012). The impact of future summer temperature on public health in Barcelona and Catalonia, Spain. International Journal of Biometeorology, 56, 1135.

Palau, R. M., M. Hürlimann, J. Pinyol, J. Moya, A. Victoriano, M. Génova and C. Puig-Polo, 2017: Recent debris flows in the Portainé catchment (Eastern Pyrenees, Spain): analysis of monitoring and field data focusing on the 2015 event. Landslides, 13, 1161-1170.

Park, S., M. Berenguer, D. Sempere-Torres, 2018: Analysis of 3-year gauge-adjusted pan-European radar rainfall accumulations. In preparation.

Pautasso, C., W. Erik, A. Rosa, 2014: REST: Advanced Research Topics and Practical

Perkins, S., & Alexander, L. (2013). On the Measurement of Heat Waves. *Journal of Climate, 26*, 4500–4517.

Poletti, M. L., Pignone, F., Rebora, N., & Silvestro, F., 2017: Probabilistic hydrological nowcasting using radar based nowcasting techniques and distributed hydrological models: application in the Mediterranean area. Geophysical Research Abstracts, 19, 14367.

Queffeulou, P., and Croizé-Fillon,D. 2014: Global altimeter SWH data set,Rep., Laboratoire d'Océanographie Spatiale, IFREMER.

Rebora, N. and Silvestro, F., 2012: PhaSt: stochastic phase-diffusion model for ensemble rainfall nowcasting. IAHS-AISH publication, 305-310.

Richardson D, Bidlot J, Ferranti L, Ghelli A, Haiden T, Hewson T, Janousek M, Prates F, Vitart F (2012). Verification Statistics and Evaluations of ECMWF Forecasts in 2011–2012. Technical Memorandum 688.

Sánchez-Benítez, A., García-Herrera, R., Barriopedro, D., Sousa, P., & Trigo, R. (2018). June 2017: The Earliest European Summer Mega-heatwave of Reanalysis Period. *Geophysical Research Letters, 45*, 1955–1962.

Siccardi, F., Boni, G., Ferraris, L., & Rudari, R., 2005: A hydrometeorological approach for probabilistic flood forecast. Journal of Geophysical Research: Atmospheres, 110(D5).

Silvestro, F., Gabellani, S., Delogu, F., Rudari, R., Boni, G., 2013. Exploiting remote sensing land surface temperature in distributed hydrological modelling: the example of the Continuum model. Hydrology and Earth System Sciences, 17, 39–62. http://dx.doi.org/10.5194/hess-17-39-2013

Thielen, J., Bartholmes, M., & de Roo, A. (2008). *The European Flood Alert System* (Vol. 5). Hydrology and Earth System Sciences Discuss, J1 - HESSD,.

Tobías, A., de Olalla, P., Linares, C., Bleda, M., Caylà, J., & Díaz, J. (2010). Short-term effects of extreme hot summer temperatures on total daily mortality in Barcelona, Spain. *International Journal of Biometeorology, 54*, 115.

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.

Vitolo C., Di Giuseppe F., D'Andrea M. (2018) Caliver: An R package for CALIbration and VERification of forest fire gridded model outputs. PLOS ONE 13(1): e0189419. https://doi.org/10.1371/journal.pone.0189419

WWRP/WGNE Joint Working Group on Forecast Verification Research. (2015, January 26). *Forecast Verification*. Retrieved April 2018, from http://www.cawcr.gov.au/projects/verification/