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

Project Title:

ANYWHERE

EnhANcing emergencY management and response to extreme WeatHER and climate Events

Subject:

Deliverable 2.2:
Preliminary report assessing the robustness and uncertainty of models to forecast weather-induced hazards and impacts

Dissemination Level:
PU Public

Delivery date: **20th December 2017**
Month: **Month 19**
Organisation name of lead contractor for this deliverable: **University of Geneva**

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

Document Information

Title	**Assessing the robustness and uncertainty of models to forecast weather-induced hazards and impacts**
Lead Author	Juan Antonio Ballesteros Cánovas (UNIGE)
Contributors	Mario Rohrer (MeteoDat) Paul Smith (ECMWF); Claudia Vitolo (ECMWF); Estíbaliz Gascón (ECMWF); Francesca Di Giuseppe(ECMWF); Claudia Di Napoli (UoR, Reading); Daniel Sempere-Torres (UPC); Henny van Lanen (WUR); Samuel Jonson Sutanto (WUR); Remko Uijlenhoet (WUR); Marc Berenguer (UPC); Shinju Park (UPC); Carles Corral (UPC); Ilona Láng (FMI); Pekka Tiainen (FMI); Tuomo Bergman (FMI); Tomás Fernández Montblanc (UNI FER); Paolo Ciavola (UNI FER); Paolo Fiorucci (CIMA); Flavio Pignone (CIMA); Peter Salamon (EC-JRC), Markus Stoffel (UNIGE)
Distribution	PUBLIC
Document Reference	Ballesteros Cánovas, J.A., Rohrer, M., Smith, P., Vitolo, C., Gascón, E., Di Giuseppe, F., (WUR); Claudia Di Napoli, Sempere-Torres, D., Van Lanen, H.A.J., Sutanto, S.J., Uijlenhoet, R., Berenguer, M., Park, S., Corral, C., Láng, I., Tiainen, P., Bergman, T., Fernández Montblanc, T., Ciavola, P., Fiorucci, P., Pignone, F., Salamon, P. and Stoffel, M. (2017): Assessing the robustness and uncertainty of models to forecast weather-induced hazards and impacts. ANYWHERE Report, Genova, Switzerland.

Document History:

Date	Revision	Prepared by	Organisation	Approved by
28.11.2017	Rev_1_0	Juan Antonio Ballesteros Canovas ; Mario Rhorer; Markus Stoffel	University of Geneva	Paul Smith, Henny Van Lanen, Paolo Ciavola, Marc Berenguer
30.11.2017	Rev_2	Juan Antonio Ballesteros Canovas	University of Geneva	Jarmo Koistinen; Markus Stoffel, Daniel Sampere, Henny Van Lanen, Marc Berenguer,
04.12.2017	Rev_3	Juan Antonio Ballesteros Canovas	University of Geneva	Markus Stoffel, Mario Rhorer
05.12.2017	Rev_4	Juan Antonio Ballesteros Canovas	University of Geneva	Henny Van Lanen, Marc Berenguer, Daniel Sempere-Torres
05.12.2017	Rev_5	Juan Antonio Ballesteros Canovas	University of Geneva	Henny Van Lanen, Marc Berenguer, Daniel Sempere-Torres
19.12.2017	Rev_6	Daniel Sempere-Torres	CRAHI-UPC	Marc Berenguer

Related Documents:

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

Summary

A preliminary assessment of uncertainty and robustness of algorithms/tools encapsulated or connected to the ANYWHERE Multi-Hazard Early Warning System is provided. The methodology is based on the identification of main sources of uncertainties and their assessment for each algorithm/tool.

Results from a specific questionnaire show the large variety of sources of uncertainty for the algorithms/tools and highlight the role of input and state variables. Results show whether algorithms/tools were tested against very extreme events, cascading effects and climate change conditions.

The results of this preliminary analysis also showed that uncertainties related to these factors are at present insufficiently tested, hence, further work is required to be able to complete a consistent analysis of this crucial aspect, with relevant consequences on the usability of the products provided by the MH-EWS.

Therefore, a new Task 2.8 has been added to WP2 to cover an extension of this preliminary study to obtain a more detailed assessment on uncertainty and robustness during the demonstration period at the ANYWHERE Pilot Sites. The purpose will be to map the suitability of algorithms/tools in (qualitative) terms for the Pilot Sites, and as far as possible for dominant pan-European physiographic settings. The uncertainty of the algorithm/tools approach will be validated by comparing impacts derived from forecasted natural hazards with observed impacts (e.g. including validation runs, sensitivity analysis). Past events will be identified to investigate impacts of coinciding and cascading effects of multi-hazards. Parallel to investigating historic and ongoing conditions impacts under a future climate (e.g. RCP8.5) will be explored to test robustness. The findings, incl. potential strengths and weaknesses, but also improvements will be documented in a new Deliverable D2.5, which will be due in M36 (May 2019).

Therefore, the performance of these models/algorithms/tools will be continuously monitored and evaluated during the demonstration period at the ANYWHERE Pilot Sites and will constitute a valuable indicator for the market uptake of the algorithms/tools realized in the framework of ANYWHERE.

For the moment, a summary of the preliminary results obtained up to now are described here.

Table of Contents

Foreword

This Deliverable, D2.2, **Preliminary report assessing the robustness and uncertainty of models to forecast weather-induced hazards and impacts** was initially conceived in the DoA as the result of the interaction of all the tasks included in the Work Package 2 of the Project: "Advanced forecasting models and tools to anticipate Weather and Climate events induced impacts", and was initially foreseen to be submitted in Month 18.

However, the investigation in the first 12 months of the project **ANYWHERE** showed that clear gaps in knowledge about uncertainty and robustness of the concerned algorithms exist. Thus, the Consortium has thought that this subject is too complex and relevant to be achieved following the initial plan and that the project would significantly benefit if the study could be extended during the duration of the demonstrations planned in the Pilot Sites of the project, where the application of the algorithms will be tested on real emergency events, and the information provided will be exhaustively tested by the operational users of the project.

From this recognition, the Consortium has proposed **to add a new Task 2.8: Robustness and uncertainty of algorithms to assess weather-related event induced impacts** specifically focused in the exhaustive analysis of the robustness and uncertainties related to the algorithms included in the Multi-Hazard Early Warning System (MH-EWS). This Task 2.8 has been proposed to run from Month 9 to Month 36 in order to take the maximum profit of the demonstrations (analysis of real events, analysis of false alarms and failures, feed-back from operators….) and also to use the data collected from the implementation of the MH-EWS to run off-line calculations on robustness and uncertainty analysis.

In consequence, it has been proposed to delay the submission of deliverable 2.2 to the end of this new Task 8, at the end of the demonstration period in Month 36.

This decision was approved in Executive Board meeting of February 2017, and then **included in proposal of amendment we planned to submit before summer 2017**. Unfortunately, this Amendment is delayed up to the end of the justification of the first reporting period due to a problem that was not possible to solve in time with the major reason justifying the amendment.

The situation was presented to the Project Officer, who expressed a concern with the idea of moving a deliverable from a reporting period to the next one. The agreement was **to split D2.2 in two deliverables:**

(i) **Keep D2.2 in Month 18 as a Preliminary report** on robustness and uncertainty of algorithms to assess weather-related event induced impacts oriented to expose the methodology (and restrict the level of dissemination to Confidential).

(ii) **Add a new D2.5 consisting on a Final report** on robustness and uncertainty of algorithms including the extended analysis performed during

the demonstration in the pilot sites (due in Month 36). This will be a Public document (Dissemination level: PU)

As this decision was just known in November 2017, the Consortium requested to delay the date of Submission of the new D2.2 to extend during several weeks the review and improvements of the document, since the authors have planned to move it to Month 36, as internally agreed by the Consortium in February 2017.

This is the reason this Deliverable is submitted later than the deadline in the initial DoA.

1 Introduction

The ANYWHERE project aims at implementing a Multi-Hazard Early Warning System (MH-EWS) at the European scale to improve the preparedness and response capacity against weather-induced hazards. The different types of hazards that the project looks at include:

- Floods, flash floods, debris flows, and landslides
- Storm surges
- Heatwaves and air quality (weather-induced health impacts)
- Weather-induced fires
- Droughts
- Convective storms, severe winds, and heavy snowfall

Calibrated and validated models/algorithms/tools are in fact powerful tools to forecast weather-induced hazards and associated impacts, and constitute the core of many early warning systems worldwide. Yet, these models/algorithms/tools are generally simplifications of complex phenomena, which include the interaction of different socio-environmental systems. The MH-EWS implemented in ANYWHERE relies on existing models/algorithms/tools that have been developed and applied in different European contexts. Identifying and quantifying the reliability of these algorithms is therefore a key issue to determine their performance to forecast natural hazards and related impacts under current and future climate conditions.

Several approaches have been used to estimate robustness and uncertainties (U&R) of algorithms/tools/models. Robustness measures based on signal-to-noise ratio and probability score have, for instance, been proposed by Knutti and Sedláček (2012) to evaluate climate model projection embedded in CMIP5. Also, changes in the direction and magnitude of model outputs were used by Tebaldi et al. (2011) and McSweeney et al. (2012) to estimate qualitatively the degree of robustness. IPCC (2007) developed a framework to evaluate the level of confidence in key research findings, based on a hierarchical assessment and the degree of consensus among experts. All these approaches have recognized the difficulty and drawbacks mostly due to the diversity of algorithms and have therefore called for a semi-qualitative assessment.

The deliverable presented here aims at providing a first assessment on uncertainties and robustness of algorithms/tools that are used to nowcast/forecast weather-induced natural hazards compiled in Work Package 2 (WP2) “Advanced forecasting models and tools to anticipate weather event induced impacts” (see details in the Deliverable D2.1) and that are implemented in the operational MH-EWS (WP3). Thus, this deliverable contributes to:

- Identifying potential sources of uncertainty by identifying and collecting basic information of each model/algorithm/tool. To this end, a questionnaire was

designed and circulated among all developers of models/algorithms/tools to be the base to carry out this identification.

- Report on uncertainties and robustness in qualitative or quantitative terms of each model/algorithm/tool based on the detailed information provided by developers.

Section 2 provides a definition of each technical term used in this report. Section 3 presents the methodology applied to estimate robustness and uncertainties of each model/algorithm/tool. Section 4 presents results from the questionnaire and Section 5 the preliminary assessment of uncertainties and robustness per model/algorithm/tool. Section 6 gives concluding remarks, and discusses the results and methodology.

After this first assessment, uncertainty and robustness of the suite of the algorithms/ tools will be exploited further, in particular at the ANYWHERE Pilot Sites. The MH-EWS is now ready to be used, which allows analysis of multi-hazards rather than single, isolated hazards, as has been the case before. Possible, cascading hazards, which often have large impacts, can therefore be investigated. Moreover, robustness of the algorithms/tools will be further tested, among others under future climate conditions.

2 Terminology

Below are the terms defined that are used in this deliverable report.

Aleatory uncertainty: aleatory uncertainty refers to the inherent uncertainty due to the intrinsic variability of the phenomena, usually expressed through probabilities. This type of uncertainty is irreducible, since the variability in the underlying variables will always exist.

Epistemological uncertainty**:** epistemic uncertainty refers to limited knowledge we may have about the system (modelled or real). This type of uncertainty is reducible.

Robustness**:** in system theory, robustness refers to the ability of a specific system to resist change without adapting its initial stable configuration.

Reliability**:** in statistics, reliability refers to the consistency of the outputs under consistent conditions.

State variables**:** State variables are a set of variables used to describe a specific dynamical system to determine its future behaviour.

Model parameter**:** combinations of (usually) constant values, which allow the models to reproduce the response of the dynamical system.

False Alarm Rate (or ratio)**:** in Early-warning Systems, the number of times a warning is issued when an event does not occur, divided by the number of times the warning is issued.

Hit rate (or ratio)**:** Number of times a warning is issued and an event occurs, divided by the number of times the warning is issued.

Receiver operating characteristic (ROC)**:** Curve that is created by plotting the hit rate against the false alarm rate at various threshold settings.

Area under ROC curve (AUC)**:** Measure for the usefulness of a warning. An AUC of 0.5 is a random predictor; an AUC of 1.0 is a perfect forecast.

Cascading effect**:** in risk assessment, cascading effect refers to a sequence of events in which each one produces the circumstances necessary for the initiation of the next.

3 Methodology to assess Uncertainty and Robustness (U&R)

This Deliverable provides an overview of the methodological assessment of Uncertainty and Robustness (U&R) of the meteorological forecast and nowcast as well as the algorithms/tools implemented in the MH-EWS to anticipate weather-induced natural hazards and impacts.

The MH-EWS consists of a series of algorithms/tools that generate probabilistic forecasts of specific weather-induced natural hazards. While probabilistic forecasting is highly useful for the preparedness and response during extreme events, yet the reliability of these algorithms/tools affects directly the quality and robustness of the forecasting. However, the large variety of the total set of algorithms/tools included in the MH-EWS, as well as the different physiographic settings where they have been tested, makes the identification of common U&R indicators highly challenging.

In this report, we have used a two-step methodology that aimed at identifying the sources of uncertainty of each model/algorithm/tool and characterizing qualitatively (and quantitative, when possible) U&R of each model/algorithm/tool.

First, we identified and recollected the possible sources of uncertainty associated to each model/tool/algorithm by circulating a questionnaire among developers. The questionnaire was circulated in the spring 2017. Outputs from this questionnaire allowed us to map the current situation and the state-of-art of algorithms/tools regarding the assessment of U&R. Specially, we aimed to identify the robustness level of these algorithms against future conditions, extreme events, cascading effects and false warning alarms. This task allowed us to determine strong points, regarding the reliability, of each model/algorithm/tool as well. The outputs have been presented and discussed among all developers (Work Package meeting held in Ferrara, Italy, June 2017).

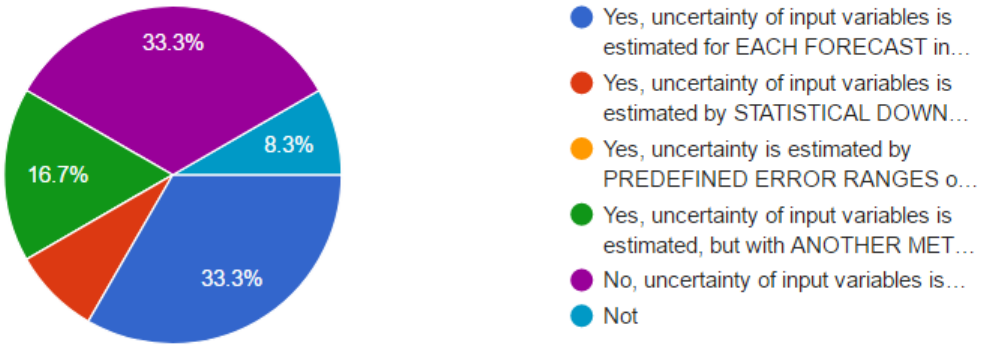
Second, an individual assessment of robustness and uncertainties of each model/algorithm/tool is provided. To this end, developers were contacted and asked for specific inputs, including a specific description of each model/algorithm/tool as well. Finally, we summarized main points of some models/algorithms/tools and their related drawbacks in the context of future climate change condition in Europe.

In a second phase, an approach will be developed and validated for each Pilot Site, by comparing impacts derived from forecasted natural hazards with observed impacts. Parallel to investigating ongoing conditions impacts under a future climate (e.g. RCP8.5 to test robustness) will be explored. If possible, outcome from a Regional Climate Model (RCM) will be used to quantify regional changes in inputs and state variables, and associated extreme events. The findings, including potential strengths and robustness, but also improvements will be documented in the new Deliverable D2.5, which is expected to be due mid 2019.

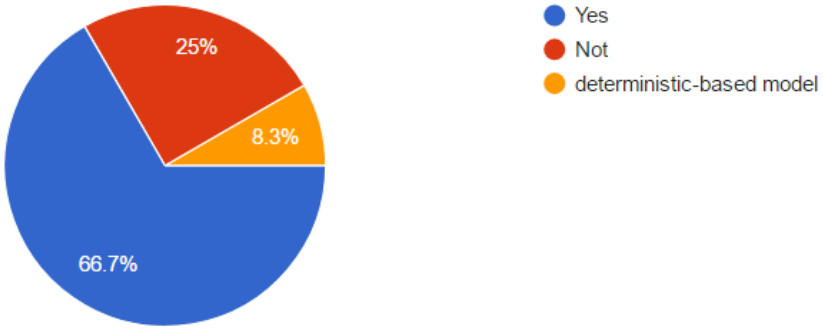
4 Results from the Questionnaire

The questionnaire was composed of 10 specific question and shared with the developers of the ANYWHERE MH-EWS algorithms in spring 2017. Overall, outputs from the questionnaire can be summarized as follows: in the 41.6 % cases, the developers did not estimate uncertainties related to input variables; although this values were, lower (25% of the cases) for uncertainties related to state variables, and parameter variability (33% of cases). Besides, the questionnaire showed that in a 16% of cases, models/algorithms/tools did not account for uncertainties related to extreme events, and on ~37% of cases the False Alarm Raito (FAR) was neither contemplated. Interestingly, none of the models/algorithms/tools account for uncertainties related to cascading effects, and only ~ 9% of cases, modelers included uncertainties due to climate change conditions. The next figures show the specific results for each of the question addressed in the questionnaire.

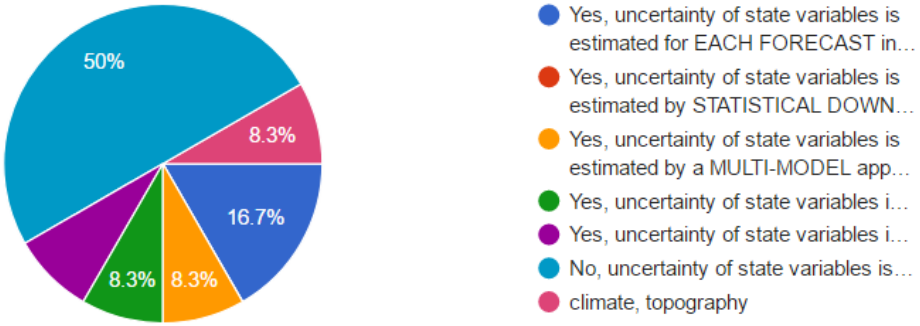
Question #1: Account for uncertainties of the input parameters: you estimate the uncertainties of the input variables of your model/tool/algorithm.



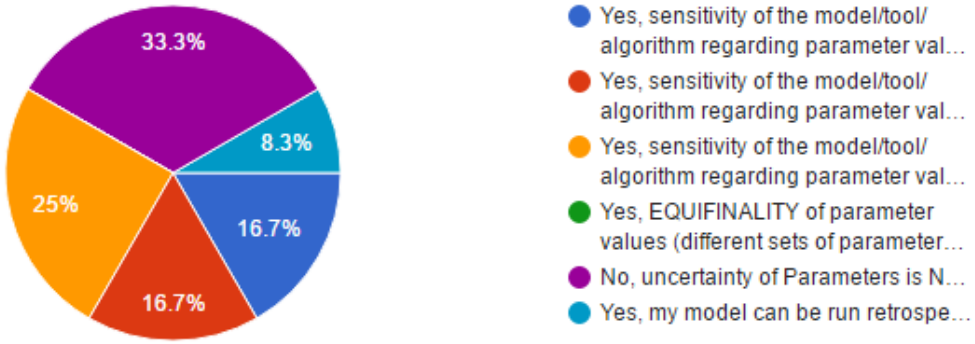
Question #2: Model sensitivity: Is your model/tool/algorithm sensitive to state variables (i.e. soil moisture; forest density; tide; etc…)



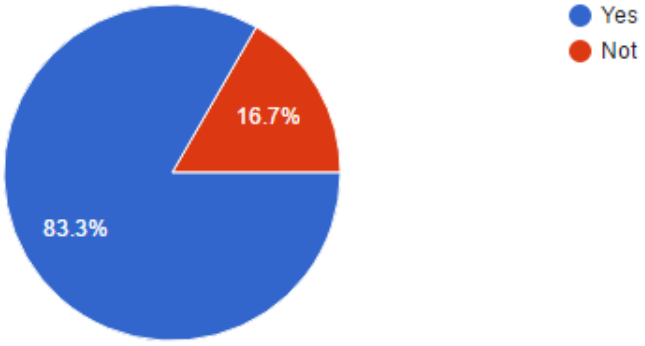
Question #3: Model sensitivity: did you account for variability in model/tool/algorithm performance to uncertainty of STATE VARIABLES?



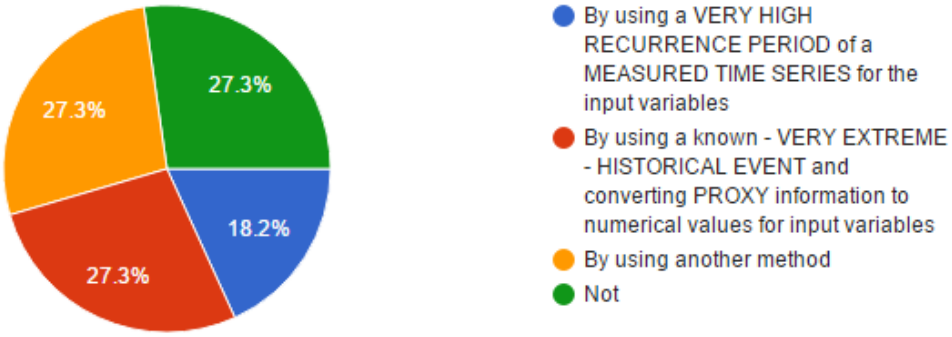
Question #4: Model variability: Did you assess model/tool/algorithm uncertainties due to parameter variability?



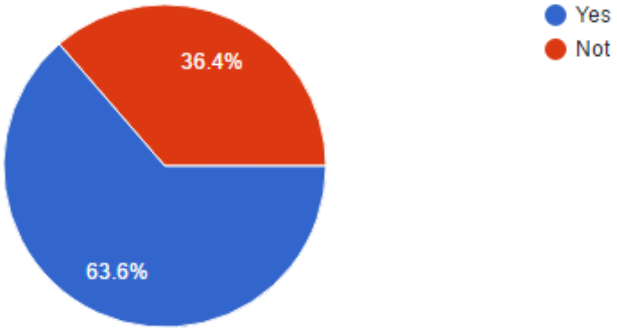
Question #5: Extreme events: Did you test your model/tool/algorithm on very extreme events



Question #6: Very extreme event: How did you test for Very Extreme Events?



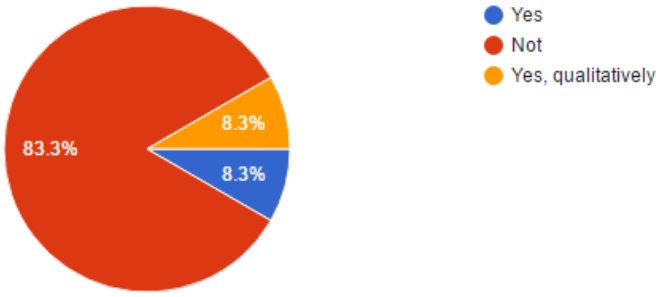
Question #7: False alarm rate: Did your model/tool/algorithm assess the False Alarm Rate (FAR)?



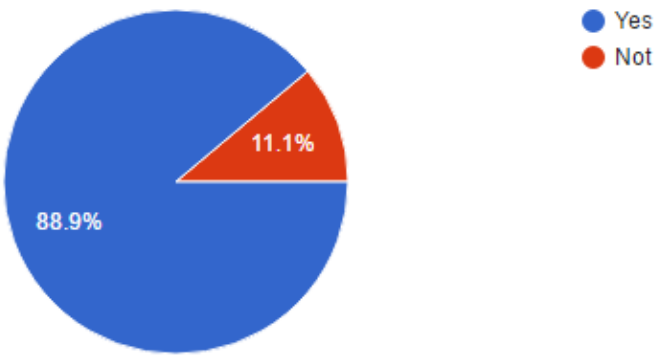
Question #8: Cascading effect: Did you account for uncertainties due to events related to cascading effects?

- 100% cases did not tested against cascading effects.

Question #9: Future Scenarios: Have you already tested your model/tool/algorithm under future Climate Change conditions?



Question #10: If NO, are you willing to test your model/tool/algorithm under future Climate Change scenarios?



5 Assessment of U&R of algorithms/tools

Since future climate is a key element for testing robustness of models/algorithms/tools, this section provides a brief description of the main impact of climate change on meteorological variables (e.g. temperature, precipitation) over Europe. Afterwards, a first discussion about the assessment of uncertainties and robustness regarding each natural hazard considered in the MH-EWS is presented.

5.1 Climate change variability in Europe and impact on the uncertainty and robustness assessment.

The recent paper published by Keuler et al., (2015) describe the climate change variability over Europe based on the outputs from the ensemble of eight regional climate models, i.e. COSMO-CLM. The assessment is based on two different greenhouse gas scenarios (RCP4.5 and RCP8.5) and four different driving GCMs (MPI-ESM-LR, HadGEM2-ES, CNRM-CM5, EC-EARTH), and covers the period from 1950-2100 at 12km grid coverage.

Although these results reveal large differences depending season and emission scenarios, some variables show changes in the same direction, which seems to be robust features. In general, surface temperature over Europe will increase. This is clearly showed in the Figure 1, where a moderate (RCP2.6) to very high (RCP8.5) increase of annual mean temperature is projected for the period 2081-2100 (relative to the period 1981-2010), especially over the Iberian Peninsula and the European Alps due to an anomalous summertime warming. As consequences, an increase of the maximum length of dry spell is likely to increase as well, severely over Iberian Peninsula under RCP 8.5 (Figure 2).

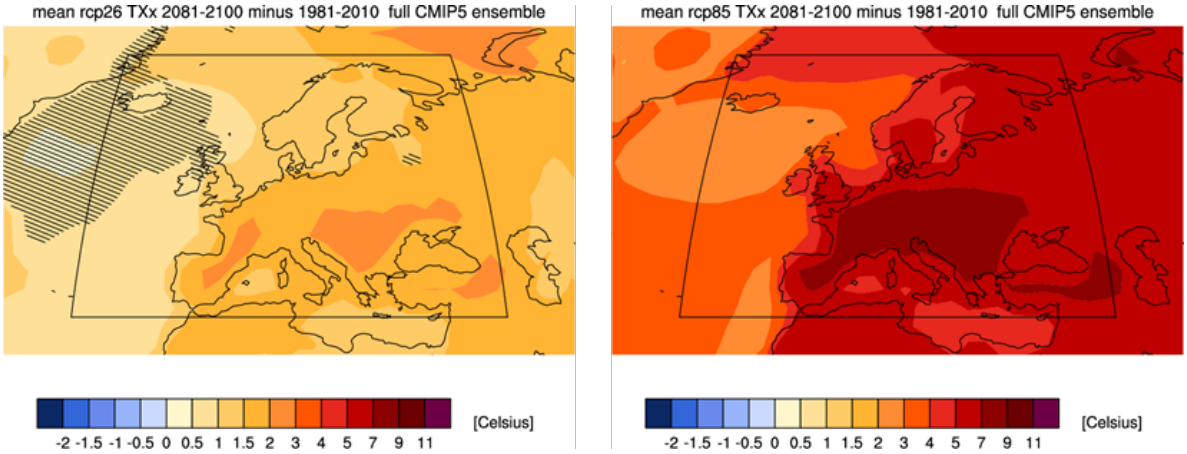


Figure 1: Mean RCP26 (left) and RCP85 (right) relative TXx (annual maximum values of daily maximum temperature) 2081-2100 minus 1981-2010 full CMIP5 ensemble. The hatching represents areas where the signal is smaller than one standard deviation of natural variability.

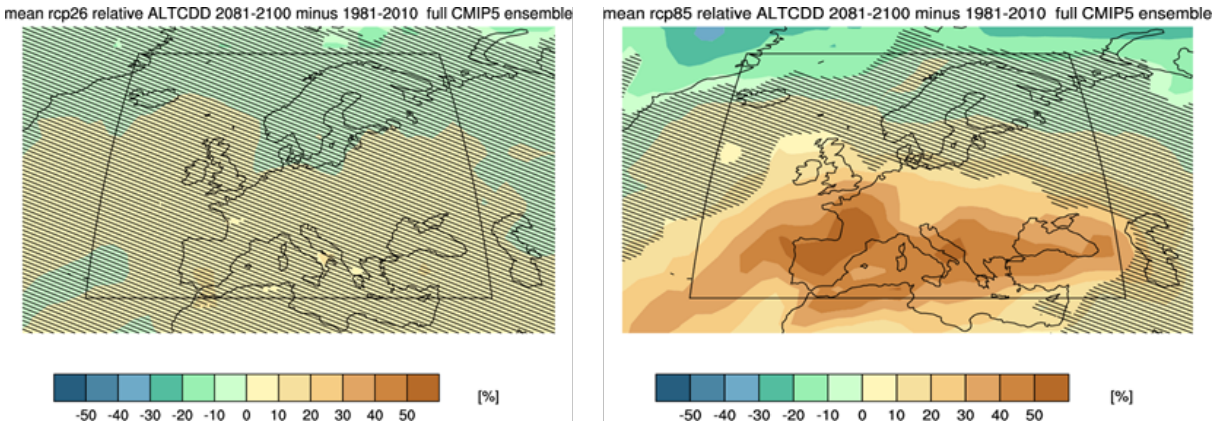


Figure 2: Mean RCP26 (left) and RCP85 (right) relative ALTCDD (maximum length of dry spell) 2081-2100 minus 1981-2010 full CMIP5 ensemble. The hatching represents areas where the signal is smaller than one standard deviation of natural variability

Regarding precipitation changes, the mean annual precipitation will slightly increase in Northern and North-eastern Europe, but decrease over Southern Europe, particularly in summer. The number of low-intensity rainfall events will decrease, while high intensity precipitation may increase, mainly as result of non-convective, large-scale precipitation episodes in winter spring and autumn. This is showed in the Figure 3.

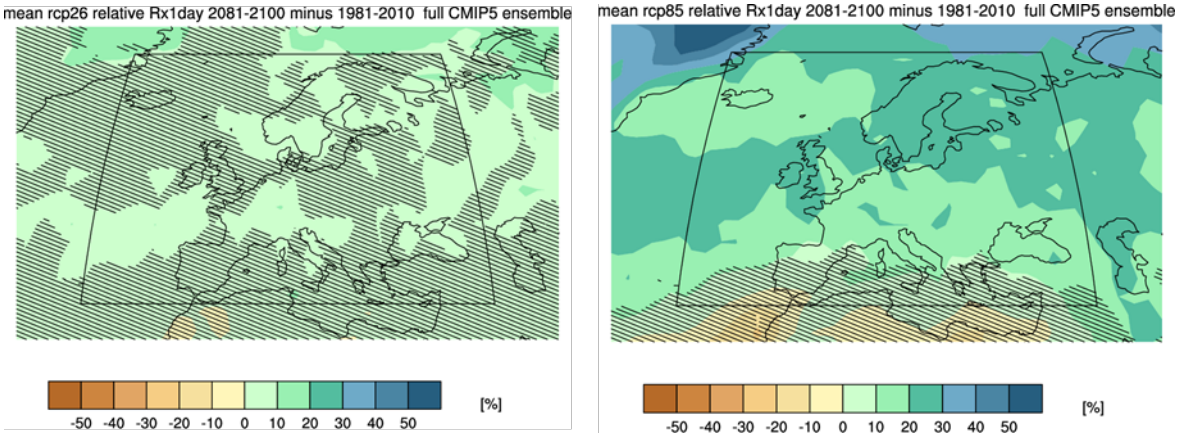


Figure 3: Mean RCP26 (left) and RCP85 (right) relative Rx (annual maximum precipitation 1-d) 2081-2100 minus 1981-2010 full CMIP5 ensemble. The hatching represents areas where the signal is smaller than one standard deviation of natural variability.

5.2 Algorithms for precipitation forecast

5.2.1 Uncertainty in precipitation forecast

Uncertainties for precipitation forecasts have been studied and verified probabilistically for precipitation type products, while the data from most probable precipitation type were considered to constitute binary observations (occurrence or non-occurrence) in the verification. This text summarizes the verification of both products, but more complete information can be found in Gascón et al. (2017).

The verification of the two new products was performed exclusively using 3-hourly observations of present weather from manned SYNOP stations in Europe. The period analysed ran from 15 October 2016 to 15 February 2017 (4-months over winter). The aim here was to assess the most recent ECMWF model cycle running over a winter period (cycle 43r1) with 3-hourly observations. The total number of stations used in this study is 1050 (Figure 4). However, not all stations are open 24 hours a day, so the nominal maximum frequency of SYNOP observations providing a current weather group during the study period varied with time of the day. The original present weather reports were classified into one of five different categories: rain (RA), snow (SN), sleet or rain and snow mixed (RASN), freezing rain (FZRA) and ice-pellets (IP) (Table 2). Wet-snow (WSN) was not considered separately due to the lack of direct observations for its verification; instead, the WSN forecasts were classified as SN. During this classification, same observation could correspond to two or three different categories (e.g. “rain or drizzle and/or snow”, with codes 68 and 69, are included in both categories, RA and RASN). Slight freezing drizzle (code figure 56) was considered “no precipitation”; however, slight continuous not freezing drizzle (code figure 51) was included in RA category (Table 1). This decision was made after many tests during the verification process, trying to avoid including extra bias due to the wrong classification of the observations (not shown). One important source of uncertainty is the height difference between model and observations, which is critical in near freezing temperatures. For this reason, SYNOP stations with an altitude difference of more than 200 m relative to the closest ENS point were removed from the verification.

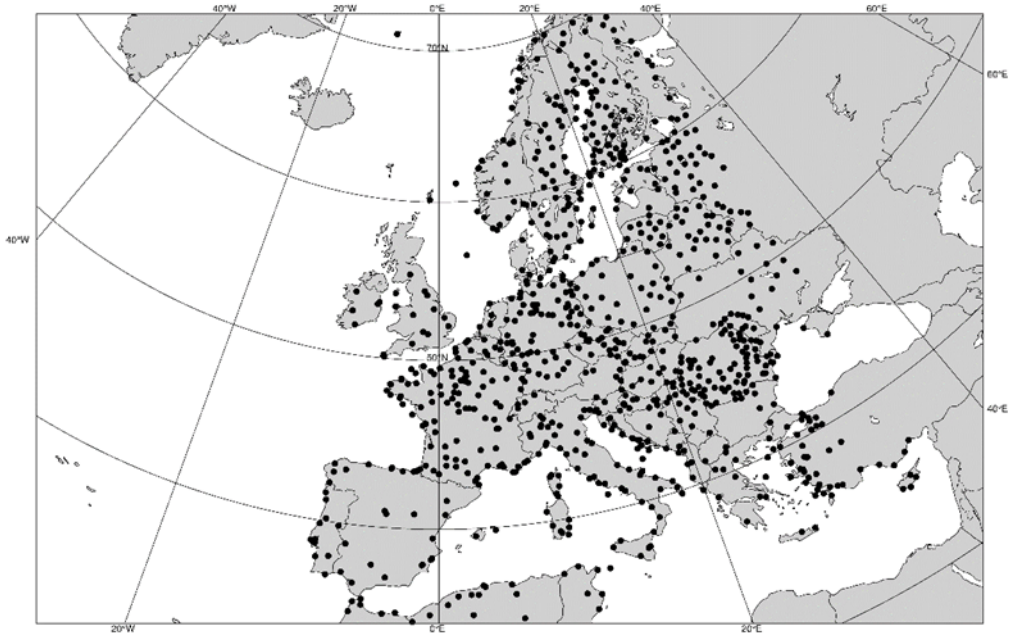


Figure 4. Map showing the SYNOP manual stations used. The verification covered the area with latitude: 33°N 35 to 71°N, longitude: 11°W to 35°E.

Table 1. Precipitation type classification from the SYNOP manual present weather code, including the number of observations of each precipitation type during the verification period (next page).

Precipitation type: 681155	Present Weather SYNOP code	Description
Rain:	51	Drizzle, not freezing, continuous slight
49892 observations	52,53,54,55	Drizzle, not freezing, intermittent or continuous, moderate or heavy
	58,59	Drizzle and rain, slight, moderate or heavy
	60,61,62,63,64,65	Rain, not freezing, intermittent or continuous, slight, moderate or heavy
	80,81,82	Rain showers, slight, moderate, heavy or violent
	87,88	Showers of snow pellets or small hail, with or without rain or rain and snow mixed, slight, moderate or heavy
	89,90,91,92,93,94	Showers of hail, with or without rain or rain and snow mixed, not associated with thunder
	95,96,97,99	Thunderstorm with or without hail, but with rain and/or snow, slight, moderate or heavy
Snow:	70,71,72,73,74,75	Intermittent or continuous fall of snowflakes, slight, moderate or heavy
34591 observations	76,77,78	Diamond dust or snow grains or isolated star-like snow crystals
	85,86	Snow showers, slight, moderate or heavy
	87,88	Showers of snow pellets or small hail, with or without rain or rain and snow mixed, slight, moderate or heavy
	89,90,93,94	Showers of hail, with or without rain or rain and snow mixed, not associated with thunder
	95,97	Thunderstorm with or without hail, but with rain and/or snow, slight, moderate or heavy
Rain and Snow:	68,69	Rain or drizzle and snow, slight, moderate or heavy
Mixed:	83,84	Showers of rain and snow mixed, slight, moderate or heavy
3362 observations	87,88	Showers of snow pellets or small hail, with or without rain or rain and snow mixed, slight, moderate or heavy
	89,90,93,94	Showers of hail, with or without rain or rain and snow mixed, not associated with thunder
	95,97	Thunderstorm with or without hail, but with rain and/or snow, slight, moderate or heavy
Freezing rain:	57	Drizzle, freezing, moderate or heavy (dense)
538 observations	66,67	Rain freezing, slight, moderate or heavy
Ice pellets:	77	Snow grains
1315	79	Ice pellets

observations		
Observed	50	Drizzle, not freezing, intermitent
precipitation but considered as “no precipitation”	56	Drizzle, freezing, slight
Rain:	51	Drizzle, not freezing, continuous slight
49892 observations	52,53,54,55	Drizzle, not freezing, intermittent or continuous, moderate or heavy
	58,59	Drizzle and rain, slight, moderate or heavy
	60,61,62,63,64,65	Rain, not freezing, intermittent or continuous, slight, moderate or heavy
	80,81,82	Rain showers, slight, moderate, heavy or violent
	87,88	Showers of snow pellets or small hail, with or without rain or rain and snow mixed, slight, moderate or heavy
	89,90,91,92,93,94	Showers of hail, with or without rain or rain and snow mixed, not associated with thunder
	95,96,97,99	Thunderstorm with or without hail, but with rain and/or snow, slight, moderate or heavy
Snow:	70,71,72,73,74,75	Intermittent or continuous fall of snowflakes, slight, moderate or heavy
34591 observations	76,77,78	Diamond dust or snow grains or isolated star-like snow crystals
	85,86	Snow showers, slight, moderate or heavy
	87,88	Showers of snow pellets or small hail, with or without rain or rain and snow mixed, slight, moderate or heavy
	89,90,93,94	Showers of hail, with or without rain or rain and snow mixed, not associated with thunder
	95,97	Thunderstorm with or without hail, but with rain and/or snow, slight, moderate or heavy
Rain and Snow:	68,69	Rain or drizzle and snow, slight, moderate or heavy
Mixed:	83,84	Showers of rain and snow mixed, slight, moderate or heavy
3362 observations	87,88	Showers of snow pellets or small hail, with or without rain or rain and snow mixed, slight, moderate or heavy
	89,90,93,94	Showers of hail, with or without rain or rain and snow mixed, not associated with thunder
	95,97	Thunderstorm with or without hail, but with rain and/or snow, slight, moderate or heavy
Freezing rain:	57	Drizzle, freezing, moderate or heavy (dense)
538 observations	66,67	Rain freezing, slight, moderate or heavy

Ice pellets:	77	Snow grains
1315 observations	79	Ice pellets
Observed	50	Drizzle, not freezing, intermittent
precipitation but considered as “no precipitation”	56	Drizzle, freezing, slight
Rain:	51	Drizzle, not freezing, continuous slight
49892 observations	52,53,54,55	Drizzle, not freezing, intermittent or continuous, moderate or heavy
	58,59	Drizzle and rain, slight, moderate or heavy
	60,61,62,63,64,65	Rain, not freezing, intermittent or continuous, slight, moderate or heavy
	80,81,82	Rain showers, slight, moderate, heavy or violent
	87,88	Showers of snow pellets or small hail, with or without rain or rain and snow mixed, slight, moderate or heavy
	89,90,91,92,93,94	Showers of hail, with or without rain or rain and snow mixed, not associated with thunder
	95,96,97,99	Thunderstorm with or without hail, but with rain and/or snow, slight, moderate or heavy
Snow:	70,71,72,73,74,75	Intermittent or continuous fall of snowflakes, slight, moderate or heavy
34591 observations	76,77,78	Diamond dust or snow grains or isolated star-like snow crystals
	85,86	Snow showers, slight, moderate or heavy
	87,88	Showers of snow pellets or small hail, with or without rain or rain and snow mixed, slight, moderate or heavy
	89,90,93,94	Showers of hail, with or without rain or rain and snow mixed, not associated with thunder
	95,97	Thunderstorm with or without hail, but with rain and/or snow, slight, moderate or heavy
Rain and Snow:	68,69	Rain or drizzle and snow, slight, moderate or heavy
Mixed:	83,84	Showers of rain and snow mixed, slight, moderate or heavy
3362 observations	87,88	Showers of snow pellets or small hail, with or without rain or rain and snow mixed, slight, moderate or heavy
	89,90,93,94	Showers of hail, with or without rain or rain and snow mixed, not associated with thunder
	95,97	Thunderstorm with or without hail, but with rain and/or snow, slight, moderate or heavy

5.2.2 Verification of Probability of Precipitation Type product

5.2.2.1 Reducing systematic bias with the observations

During the development of the Probability of Precipitation Type product, we described a methodology for rate-related frequency bias correction for the precipitation type variable, defining a Rmin (minimum precipitation rate to consider precipitation or no precipitation) for each precipitation type. The target of this procedure was to make the total frequency of occurrence of each precipitation type, within forecasts, over all the observation sites, equal the observed frequency of occurrence at those sites (i.e. frequency bias = 1).

In order to evaluate the advantages of post-processing using Rmin in probabilistic forecasts of precipitation type, reliability diagrams for each precipitation type with different thresholds were constructed. Reliability diagrams (Murphy and Winkler 1977; Wilks 1995) compare the forecast probabilities against the frequency of an event occurrence and therefore measure how closely the forecast probabilities of an event corresponds to the actual chance of the event occurring. In this section, two different lead times for evaluating performance are again considered (24-48 h and 96-120 h), with two different precipitation rate thresholds (see Figure 5). The black solid diagonal line represents perfect reliability. RA forecasts are reasonably reliable for both lead times and Rmin settings (Figure 5a), though the larger Rmin (blue), gives better results throughout. The SN forecasts are also reasonably reliable (Figure 5b), but if the larger Rmin setting were used, and probabilities were low (10-30% say), too many events would be missed. FZRA (Figure 5c) and RASN (Figure 5d) forecasts are not good but show some limited skill, though sample size seems insufficient to highlight the benefits of the recommended Rmin values. Also of note is the fact that high probability forecasts of either, though rarely occurring, are generally far too confident. This is a typical characteristic of reliability diagrams for parameters that are generally not well predicted. Finally, Figure 14e shows that probabilistic forecasts for IP cannot be relied upon.

Clearly, sample size affects the above results (Table 1). Frequencies of IP and FZRA are very low compared to, for example, RA. On the other hand, from the perspective of severe winter weather prediction, it is somewhat encouraging that in spite of this FZRA forecasts, at least, do have some reliability at day 2. This is probably because there is more spatial continuity/extent in FZRA during a FZRA event than there would be for IP during an IP event.

One disadvantage of post-processing in general is that there is always a need to re-calibrate each time a related significant change is made in a new model cycle. However, at ECMWF experimental runs covering many winter months are always carried out in advance of the release of a new cycle and a verification tool has been automatized to re-calibrate the products in case the biases change significantly. This will allow the re-calibration of Rmin to also be done in advance, and as illustrated above 4 months of re-runs should be sufficient for this purpose. It is highly probable that the BIAS varies seasonally (e.g. we would expect a larger rate threshold for RA in summer); however, this new tool presents its main use in the winter season, so we have prioritized the correction of this period of the year.

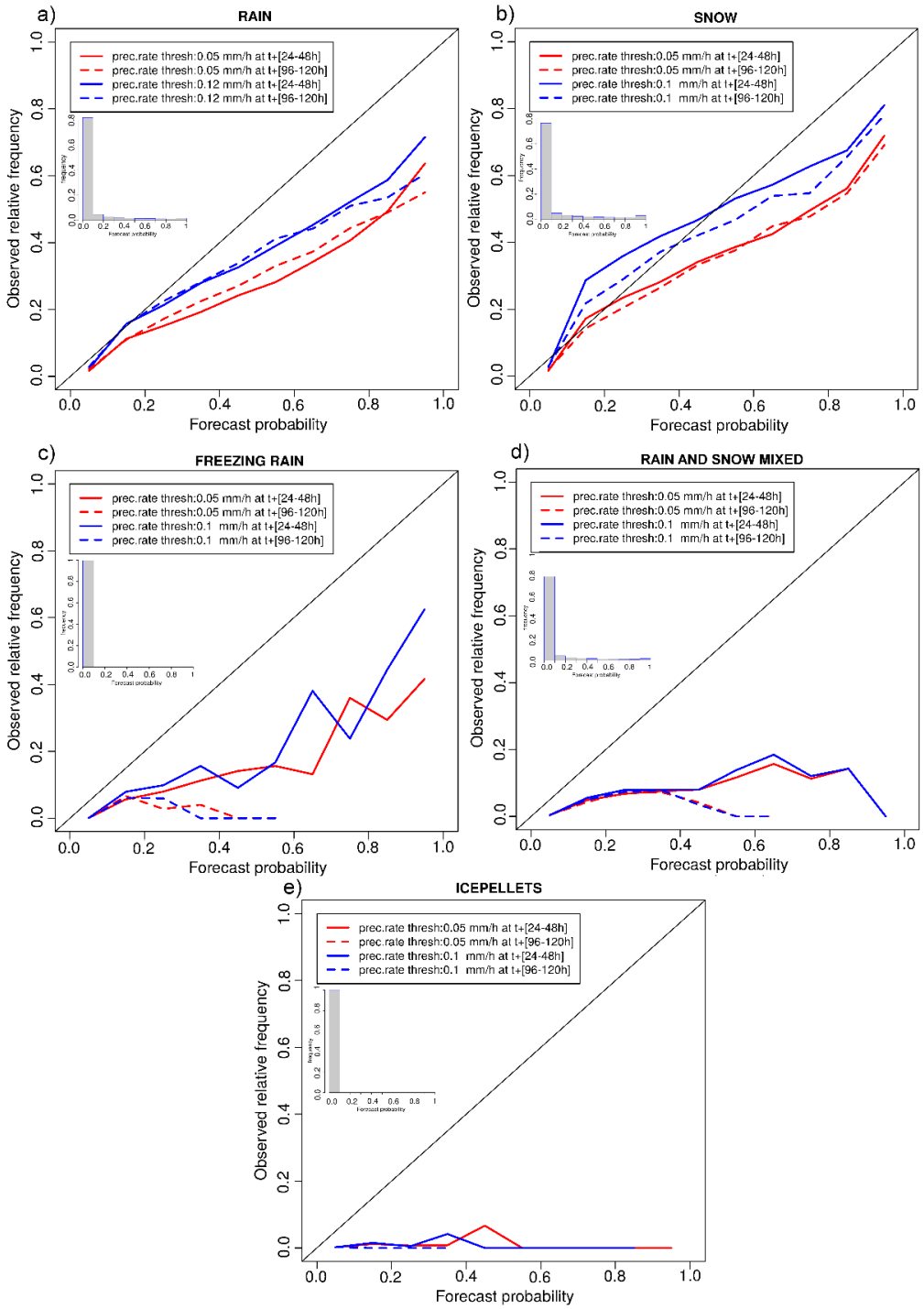


Figure 5. Reliability diagrams for ENS forecasts at 24-48 h (solid lines) and 96-120 h (dashed lines) lead times for 0.05 mm h−1 precipitation rate threshold (red lines), 0.12 mm h−1 (blue lines in a) and 0.1 mm h−1 (blue lines in b,c,d and e) for (a) rain, (b) snow, (c) freezing rain, (d) rain and snow mixed and (e) ice pellets. The inset histograms denote frequency of forecast usage of each probability bin for the 0.05 mm h−1 precipitation rate threshold.

5.2.2.2 ROC curves

The relative operating characteristic (ROC) diagram (Mason 1982) is widely used to evaluate the quality of probabilistic forecasts (Stanski et al. 1989; Buizza and Palmer 1998; Mason and Graham 1999). It plots the hit rate (H) against the false alarm rate (F), based on 2, for different probability thresholds. The main diagonal corresponds to random forecasts (H=F), and the area under the ROC curve (AUC) (Hanley and McNeil 1982) is taken as a measure of skill, with values between 0.5 (random forecast) and 1 (perfect forecast). For the verification of each precipitation type, we first apply the Rmin filter for each precipitation type, as discussed above. Following the filtering, the verification is performed without taking into account the intensity of the precipitation, only the probabilities of occurrence. Although ROC curves do not of themselves provide any measure of reliability, in our case we have tried to maximise reliability using the Rmin thresholds. However use of this approach does not mean that we will get perfect reliability at every probability threshold.

ROC curves for each precipitation type and at seven different lead times are shown in Figure 6, wherein probability thresholds were assigned at 2% intervals (although labels were added only at 10% intervals, for the shortest range forecasts, shown in red). The AUC score for each category and each time step is indicated in the lower-right corner of each diagram. Looking at ROC curves for RA (Figure 6a) and for SN (Figure 6b), they seem quite similar, and however, RA is a bit more skilfully predicted than SN. The differences between RA and SN ROC plots are clearer in the AUC values (grey boxes). AUC values for RA precipitation (Figure 6a) are between 0.90 at 0-24 h to 0.81 at 144-168 h lead time. In the SN case (Figure 6b), these values range from 0.87 to 0.83 at the same lead times. From day 5 to day 7 the RA AUC is slightly lower than the SN AUC, due to the fact that the F is greater for RA than for SN at these lead times. As in the analysis of the reliability (Figure 5), ROC curves for FZRA (Figure 15c) and RASN (Figure 6d) are quite similar, probably due to their lower frequencies in the study sample. For FZRA and RASN the first day exhibits slightly less skill in the ROC curves comparing with the second day. The AUC index for FZRA varies between 0.72 at 0-24 h to 0.59 at 144-168 h, indicating slight skill, especially at earlier lead times. RASN (Figure 3d) shows similar overall skill to FZRA, however it is slightly worse than FZRA at shorter lead times and better at longer lead times. In fact, the skill in RASN forecasts using this metric does not vary much with lead time. Finally, the ability to predict IP is almost negligible, with all curves close to the diagonal.

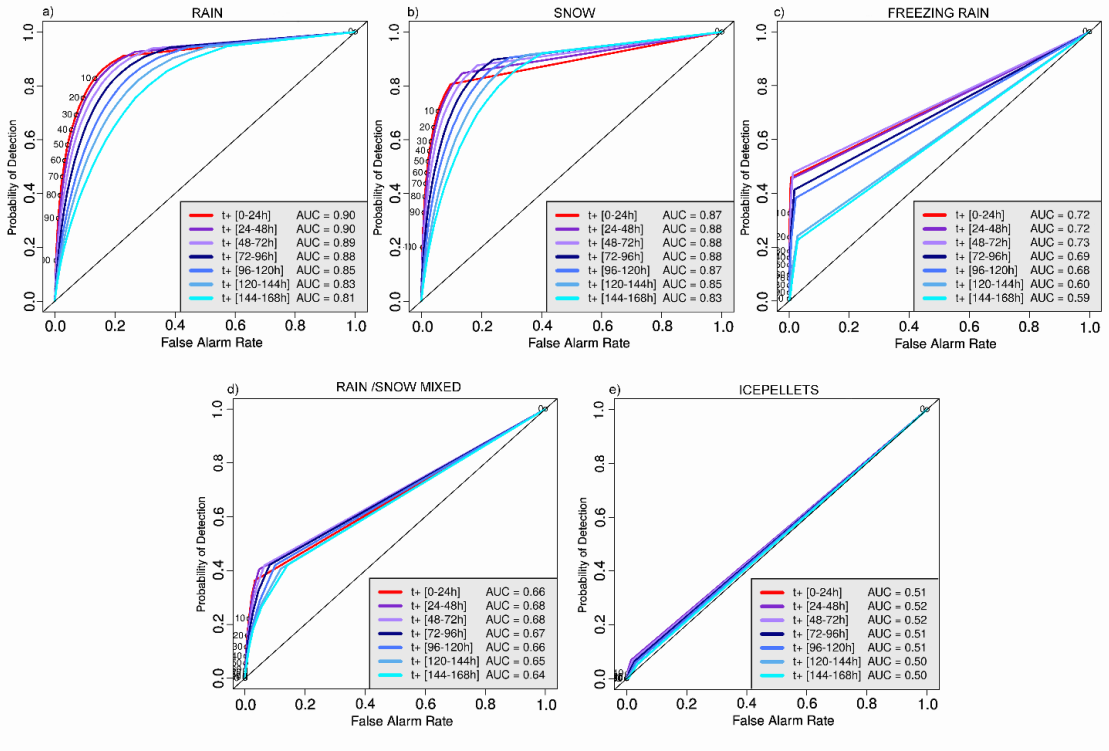


Figure 6. ROC curves at different lead times, up to day 7, for (a) rain, (b) snow, (c) freezing rain, (d) rain and 44 snow mixed and (e) ice pellets. The curves are the plots of Hit Rate vs False Alarm Rate for each decision threshold (2% interval used). Labels, at 10% intervals, are shown for the day 1 forecasts only (in red). The 45◦ angle black line represents no skill. The areas Under each Curve (AUC) for each lead time are shown in the grey boxes.

5.2.3 Verification of most probable precipitation type

The most probable precipitation type product was verified as a dichotomous (yes-no) forecast and it was applied only for areas where the total precipitation probability is > 50%. Performance diagrams (Roebber 2009) can relate four verification indices in the same plot: hit rate (H), success ratio (SR), frequency bias (FB), and critical success index (CSI; also known as threat score). This diagram is similar to the Taylor diagram (Taylor, 2001) but useful for dichotomous (yes-no) forecasts. Based on a 2x2 contingency table (Table 2), these scores are defined as:

Table 2. Standard 2x2 contingency table for dichotomous forecasts

	Event observed YES	Event observed NOT
Event (YES)	A (hits)	B (false alarms)
Forecast (NOT)	C (misses)	D (correct negatives)

$$H = \frac{A}{A+C} \qquad (1)$$

$$bias = \frac{A+B}{A+C} \qquad (2)$$

$$CIS = \frac{A}{A+B+C} \qquad (3)$$

$$SR = \frac{A}{A+B} = 1 - FA \qquad (4)$$

where False Alarm Ratio (FAR) is:

$$FAR = \frac{B}{A+B} \qquad (5)$$

These indices are mathematically related and the geometrical representation in a single diagram allows accuracy, bias, reliability and skill to be simultaneously visualized. This performance diagram for all precipitation types at each lead time is shown in Figure 7. Dashed lines represent bias scores with labels on the outward extension of the line, and labelled solid contours are CSI. Green dots correspond to RA, blue dots to SN, red to FZRA, turquoise to RASN and orange indicates IP. The different dot sizes represent different lead times, so the smaller the point, the longer the lead time. In the original conceptualization of this diagram a perfect forecast would lie in the upper right corner; however, this is a post-processed product where we obtained the most probable precipitation type, thereby eliminating the possibilities of other precipitation types, so the verification results are in effect portrayed for the product itself, and not specifically for the precipitation type variable in the ENS. For RA and SN (Figure 7), the earliest lead times are clustered toward the centre of the diagram, close to the bias=1 line, especially RA at 0-24 h lead time (the biggest green dot) with maximum H between 0.5 and 0.6 (and similar result for SR). For the same precipitation types, values of CSI between 0.3 and 0.4 are observed, decreasing to 0.1 as we move on to day 6 forecasts. As seen in the probability of precipitation type ROC curves (Figure 6), the skill for FZRA and RASN on the most probable precipitation type product is low, but there is still some predictability. In this case, the H is lower (values not higher than 0.2) than with the probability product (more than 0.4 in Figure 6c). Finally, the forecast skill of this product is minimal for RASN and completely negligible for IP.

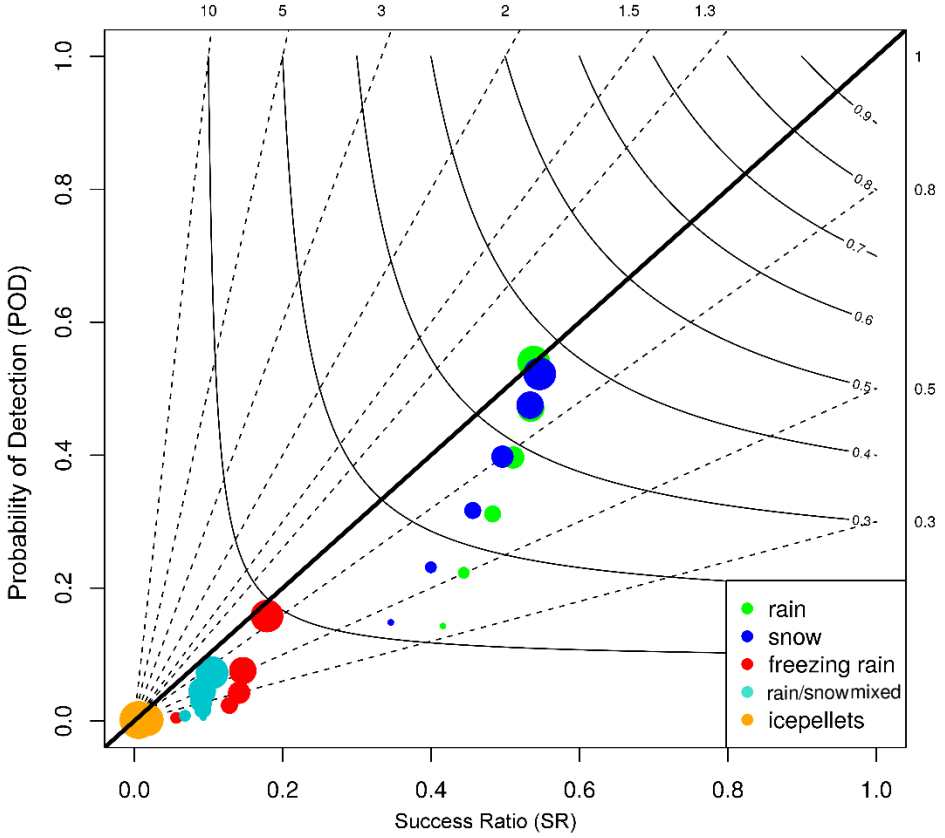


Figure 7. Performance diagram for the most probable precipitation type for each type of precipitation and for multiple lead times. Labelled solid contours represent CSI (Critical Success Index) and dashed lines are bias score with labels on the outward extension of the line. Different sizes of the points indicate the six different lead times (the bigger the size, the shorter the lead time, from 0-24 h to 144-168 h).

Because FZRA, RASN and IP are usually transient mixed precipitation phases, identification on the most probable precipitation type will tend to correspond, on average, to lower probabilities than one sees on average for RA and SN on that product. For similar reasons, as we go to longer lead times the frequency with which one sees these types on the map product reduces very rapidly (note that on Figure 4 the nominal bias for this product reduces more rapidly with lead time for mixed phase types than it does for RA and SN). Although this verification diagram does not incorporate the verification metrics which are most strongly affected by base rate dependency or the no-occurrence cases, such as percent correct (PC) and F, it still uses H, SR, and CSI, which are all potentially affected by it. When an event becomes rarer, these quantities or indexes tend towards 0 because the entries in the contingency table tend to zero at different rates. One way to solve this issue can be the computation of Symmetric Extremal Dependency Index (SEDI) which has many beneficial properties which are not present in most other verification measures used with rare events (Ferro and Stephenson 2011), and it is based in the F and H indexes. The score is defined as:

$$SEDI = \frac{logF - logH - log(1-F) + log(1-H)}{logF + logH + log(1-F) + log(1-H)} \quad (6)$$

Table 3 shows SEDI index for the three rare events studied in this paper (FZRA, IP and RASN) at different lead times. This index gives a better idea of the skill for these three precipitation types in the most probable precipitation type product than does the performance diagram, where the values of the indices were a bit too small to usefully compare at different lead times. FZRA shows better skill in shorter lead times with a maximum value of 0.61, which progressively reduces with lead time, reaching zero at 120-144 h and 144-168 h. RASN is more stable over lead times, keeping always low values that vary from 0.38 at 0-24 h to 0.19 at 144-168 h lead times. Finally, ice pellet forecasts have no skill, confirming earlier results from the probability of precipitation type product verification.

Table 3. Symmetric extremal dependence index (SEDI) for FZRA, IP and RASN at different lead times.

	0-24h	24-48h	48-72h	72-96h	96-120h	120-144h	144-168h
Freezing rain	0.61	0.51	0.46	0.41	0.29	0	0
Ice pellets	0.08	0.14	0	0	0	0	0
Rain/snow mixed	0.38	0.33	0.30	0.27	0.22	0.22	0.19

5.3 Algorithms from the European Flood Awareness System (EFAS)

European Flood Awareness System (EFAS) provides forecasts about riverine flows, riverine floods and flash floods to the MH-EWS.

The EFAS forecast performance is continually monitored in terms: (i) counting the number of notification alerts sent out (alerts are counted as hits or false alarms in comparison with observed floods) and (ii) skill scores, such as bias, Nash-Sutcliffe efficiency and continuous ranked probability scores (CRPS). Often the system is evaluated against its own climatology, which is the water balance run of the hydrological model LISFLOOD (Pappenberger et al., 2015a). Performance results are published in a bimonthly bulletin (https://www.efas.eu/efas-bulletins.html) and the scientific literature (e.g. Alfieri et al. 2014; Pappenberger et al. 2016). New user focused scores are developed as needed (e.g. Cloke and Pappenberger 2008; Pappenberger et al. 2008). The performance of the model is steadily increasing, however there are inter-annual variations.

Products based on two EFAS forecast types are utilised within the MH-EWS. The prediction of riverine floods is based upon the output of a pan-European hydrological model driven by Numerical Weather Prediction (NWP) forecasts. Flash flood

guidance results from the ERIC product, which considers the return periods of catchment wide surface runoff over different time scales. Further details can be found in the deliverable D2.1 and D2.3 of ANYWHERE project.

5.3.1 Riverine forecast robustness

The quality of the riverine forecasts within EFAS are limited by two key factors:

1. The ability of the hydrological model to represent the response of the catchments to meteorological forcing.
2. The accuracy and reliability of the predictions of the meteorological forcing.

The later of these is discussed in more detail. Below the quality of the underlying hydrological model and resulting forecasts is discussed

5.3.2 Hydrological model

The hydrological model used within EFAS is LISFLOOD a spatially-distributed hydrological rainfall-runoff model developed at the European Commission Joint Research Centre (Bartholmes et al. 2008, Knijff et al. 2010, Thielen et al. 2009a). As with most conceptual hydrological models the quality of the model performance is dependent upon the calibration of the model against observed data.

A calibration exercise completed in 2013 (Zajac et al. 2013) produced Europe wide parameter maps based on the estimation of parameter values for 693 catchments. For all catchments set of 9 parameters that control snowmelt, infiltration, preferential bypass flow through the soil matrix, percolation to the lower ground water zone, percolation to deeper groundwater zones, residence times in the soil and subsurface reservoirs and river routing, were estimated by calibrating the model against historical records of river discharge. An additional four parameters related to the represent of reservoirs were also calibrated in 34 catchments.

Figure 8 shows Nash-Sutcliffe efficiency (NSE) of the calibrated LISFLOOD model for the calibration (01-Jan-1994 to 31-Dec-2002) and validation (01-Jan-2003 to 31-Dec-2012) time periods. For the calibration period LISFLOOD is shown to have explanatory power for 90% of the catchments (NSE>0). In 32% of the catchments, LISFLOOD explains over three quarters of the variance of the observed series. Visual and numeric comparison of the calibration and validation periods show a broadly similar performance perhaps indicating some robustness to the model changes over changes in the decadal scale.

While there is overall good agreement between the observed and simulated flow statistics, large discrepancies do occur at a small number of stations, particularly in the Iberian Peninsula and on the Baltic coasts. These may be attributed to a number of sources such as errors in meteorological forcing, the spatial interpolation of meteorological data, as well as to shortcomings in the hydrological model, its static input and the calibration of its parameters. Some of the differences may also be due to those man-made modifications of flow regimes present in many catchments, but which are not fully accounted for in the hydrological model.

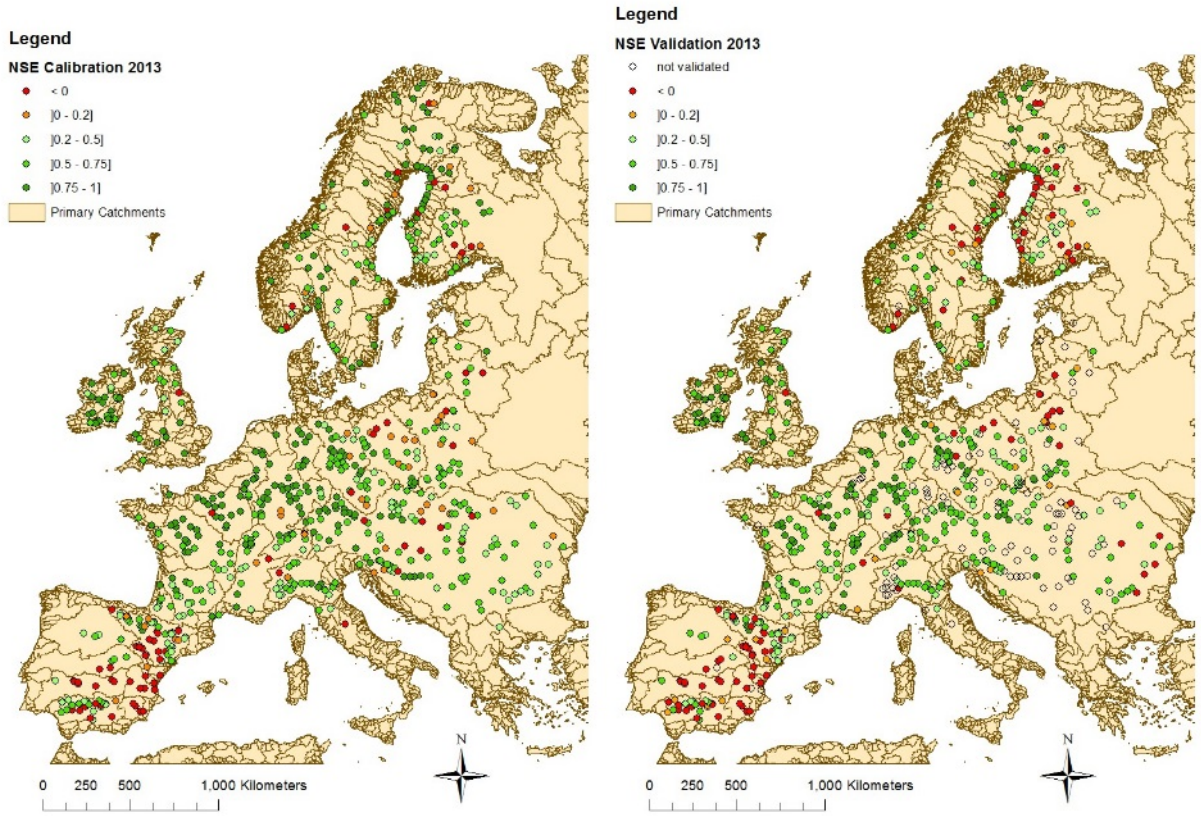


Figure 8: The Nash-Suttcliffe efficiency of LISFLOOD at the 693 sites for the calibration (left) and validation (right) periods (Smith et al 2016)

5.3.3 EFAS forecast Performance

The forecast performance of EFAS is continually monitored in two ways. The first is assessing the quality of the flood alerts watches sent to EFAS partners. Though these are not available through the MH-EWS they give an indication as to the potential to define successful warning rules from the forecast data. The second monitoring is in terms of skill scores such as bias, Nash-Sutcliffe efficiency and continuous ranked probability scores (CRPS).

As far as possible, the flood alerts are counted as hits or false alarms in comparison with observed floods, otherwise they are assigned as unknown. Figure 9 shows a large inter-annual variation, and that 2013 and 2014 stand out as having a greater number of warnings. The main reason for this are two major flood events the central European floods of 2013 and the Balkan floods of 2014.

The observed occurrence of a flood in Figure 9 was extracted from the International Disaster Database (ref: http://www.emdat.be/database, accessed 10 June 2015). There are clear trends in the data. It appears the system has increased in activity over the years when comparing the number of reported events with warnings and alerts. This could be due to the fact that the number of EFAS members have grown over the years, or that due to changes in the criteria for issuing alerts and watches more are being issued. However, there is a much higher correlation between the number of affected people and the number of issued flood alerts (0.89), than with the

number of events (0.65). This is an effect of how an event is classified in the database. The number of people affected is a better measure of the total extent of the flood, and this is what is reflected in increase in the number of alerts.

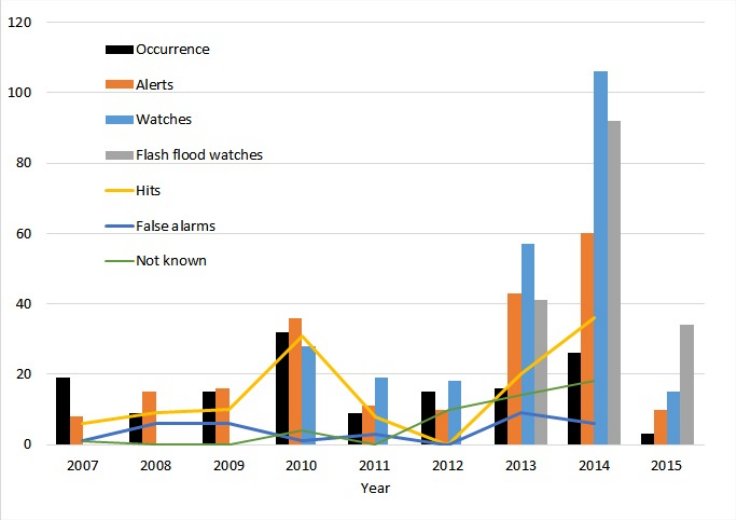


Figure 9: Number of Flood alerts, watches and flash flood watches sent since 2007. The lines show the verified hits and false alarms over the first years of EFAS.

The riverine forecasts issued by the system is evaluated against its own climatology which is the water balance run of LISFLOOD (Pappenberger et al. 2015a) using a number of skill scores. The performance of the model is steadily increasing, driven mainly by improvements to the meteorological forecasts. Figure 10 shows a degree to inter-annual variation with drops in performance often related to periods such as the winter of 2015, in which the meteorological situation was difficult to predict for all weather centres.

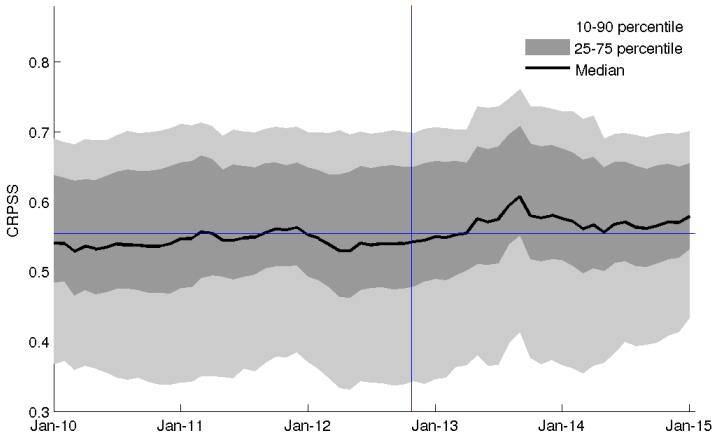


Figure 10: CRPSS of EFAS driven by ECMWF ENS over the period 1 January 2010 – 30 April 2015 as an annual running mean. The lighter grey areas show the 10-90th percentiles, and the darker grey 25-75th percentiles. The results are shown for all the areas larger than 4000 km^2.

5.3.4 Robustness of ERIC as indicator of flash flood potential

The ERIC product within EFAS is an indicator of flash flood potential. As an indicator verification and robustness testing are best carried out against reported events since an equivalent observed variable, such as discharge, is not available.

This section outlines some specific case studies where EFAS flash flood warnings from the ERIC indices were compared to records of flood events. Observed flood locations for each event are collated from reporting in the media, EMDAT and the European Severe Weather Dataset (ESWD). An ERIC alert was classified as being successfully validated if it was within 100 km, the same sub-catchment and the same day as the nearest observed reporting point. However, it should be noted that much uncertainty is associated with the 'observed' reporting locations:

- Reporting can be location selective and centred around larger settlements – therefore it cannot be stated with confidence that locations with no observed flooding definitely did not flood
- The dates of flooding reported in the media can be vague and often can only be used to narrow down to a window of a few days
- ESWD reports relate the heavy rain and not directly to flooding, the raw data were filtered to only include those which explicitly mentioned flooding however this could still relate to small-scale pluvial events for which EFAS is not designed to detect
- Since reporting is biased towards larger settlements, these points may relate to areas where the upstream catchment area >5,000 km^2 hence no ERIC warning points will be generated. Despite this EFAS flash flooding warning points may still have been generated for smaller catchments in the area.

Accounting for these uncertainties mean that it is not possible to definitively calculate the number of false alarms generated by ERIC, instead these will be referred to as 'unvalidated points'. The number of 'unvalidated points' will still be used to calculate the threat score but its reliability should be borne in the context of the above uncertainties.

5.3.5 Examples of case studies

Table 4: Summary Performance of the ERIC indicator for the considered events.

Case Study	Correct Hits	Unvalidated Points	Missed Observations	Hit Ratio	Threat Score
Central Europe: May/June 2013	212	49	121	0.64	0.56
Southern Europe: Jan/Feb 2015	214	107	22	0.91	0.62
Balkan: May 2014	450	57	97	0.82	0.75
France & Italy: October 2014	81	57	3	0.96	0.58

Table 4 summarises the performance of ERIC over the case studies. In general, ERIC performs well identifying a high proportion of the reported flood events, despite the variety of climatology, topography and land use represented by the case study sites. In the following sections a narrative description of the forecast evolution and possible causes for the missed observations or unvalidated points given for each event.

5.3.5.1 Central Europe: May/June 2013

Flash Flooding struck the central European area between the end of May and the beginning of June 2013, particularly affecting Germany, Austria, Switzerland and the Czech Republic. The precipitation which fell was not of an especially high intensity, however the catchments were highly saturated from preceding precipitation events hence leading to the possibility of flash floods being generated.

From the analysis between 30th May 2013 to 3rd June 2013 ERIC generated the majority of the warning points on the 2nd and 3rd of June, most have a persistence of 12 hours but several have a greater value of 24 and 36 hours. For the first two dates ERIC produced points in eastern Czech Republic and upper Austria on 31st May, and Vorarlberg, Austria, the Zurich and Aargau cantons of Switzerland and western Baden-Wurttemberg, Germany on 1st June.

Investigating the spatial distribution of the EPIC warning points for each day highlights the possible unreliability of the observation points on 31st May and 1st June. During both these days there are many 'observations' in the lower reaches of German rivers, but given that the most intense precipitation only fell during the 31st May in the Alps, it is unlikely that the flood peaks had been able to reach this far downstream. The bulk of these observation points were derived from EM-DAT

reporting, which vaguely states that the flooding occurred from 28th May to the 6th June.

By 2nd June 2013 flooding becomes more widespread, ERIC generates warning points along the Salzach and upper Inn rivers along the German-Austrian border as well as further downstream, for example in the Central Bohemia and Liberac regions of the Czech Republic and in Saxony-Anhalt, Germany in the upper catchment of the Elbe river. A similar pattern is also evident on the next day, 3rd June 2013. Regarding the missed observation points on these two days, it can be seen that the most northerly points, those which are furthest downstream, never have a matching EFAS warning point. Such observations occur along the lower Elbe river where the upstream area vastly exceeds the upper limit for EFAS flash flood warning point generation, hence these observations could be excluded from the analysis.

5.3.5.2 Southern Europe: January/February 2015

Between 30th January and the 2nd February 2015 a series of intense precipitation events occurred across southern Europe in countries such as Spain, Italy, Greece, Albania and Montenegro. Three deaths were reported in Bulgaria, in Pais Vasco, Spain roads and buildings were flooded and one death was recorded in Pamplona. In Albania and Greece villages were flooded, in the latter the historic Plaka bridge was destroyed. Some of the forecasted precipitation was due to fall in the mountainous regions of Albania and Montenegro thus raising the question about whether this would fall as rain or snow. If the former, then there was a possible flood risk, if the latter then false alerts may have been generated.

Most of the unvalidated ERIC warning points occur on 31st January and 1st February 2015. On the 31st January these false alerts mostly occur in Greece and Italy, however on the next day similarly distributed warning points are successfully validated, hence it is possible that the dates of the observation points do not extend to the correct range but it was not possible to verify this.

5.3.5.3 Balkans: May 2014

Between 13th – 16th May 2014 severe rainfall over the Balkans resulted in flooding, especially in Serbia and Bosnia-Herzegovina. In some areas over 200 mm of rain fell in 3 days onto ground which was already saturated. This represents the limit of what may be considered a flash flood due to the longer duration and lower intensity of precipitation. However, it does represent an example of how the flash flood indicators react to a more conventional riverine flood.

The observation points in this case used to compute the performance (Table 4: Summary Performance of the ERIC indicator for the considered events.

Case Study	Correct Hits	Unvalidated Points	Missed Observations	Hit Ratio	Threat Score
Central Europe: May/June 2013	212	49	121	0.64	0.56
Southern Europe: Jan/Feb 2015	214	107	22	0.91	0.62
Balkan: May 2014	450	57	97	0.82	0.75
France & Italy: October 2014	81	57	3	0.96	0.58

) were mostly taken from EMDAT and media reports where there was no reliable date information, so it was assumed that these locations were flooded on each of the four analysis days. This could lead to the over-estimation of correct hits and hence the hit ratio and threat score. In ERIC this problem could be exacerbated due to the production of a larger number of warning points which results in a larger number of hits. Thus, this highlights the need to be careful with the statistical analysis of the ERIC indicator values in the absence of reliable observation data.

Like in the previous case studies, ERIC produces a larger number of successfully validated warning points. Most of these points are forecast for the 15th and 16th of May, they have a persistence of between 12-24 hours (i.e. 1-2 forecasts) but with some points at each time step having longer persistence. For example at 15/5/2014 00:00 there are 6 points with a persistence greater than 6 forecasts. This may indicate a potential for ERIC to perform well event for more riverine floods in smaller basins.

5.3.5.4 France and Italy: October 2014

Between the 3rd – 4th November 2014 heavy precipitation fell in central-eastern and south-eastern areas of France before moving eastwards to northern Italy. This resulted in flooding in the Burgundy, Ardeche and Rhone-Alpes areas of France and in the Cinque Terre area of Italy. This case study is potentially interesting because flood alerts were issued for flooding on the 5th and 6th November in the Po valley west of Milan and the Adige valley in Trentino. However, the precipitation in the upstream mountains fell as snow meaning that the flood alerts were not realised. Hence, this could be a test of how ERIC deals with flood alerts when snowfall may occur, since this index accounts for the distinction between rainfall and snowfall.

On 5/11/2014 ERIC, warning points are clustered along the Saone river between Dijon and Lyon which corresponds to the heavy precipitation which fell in the latter half of the previous day and the early hours of the 5th November. There are clusters of unvalidated warnings between Cannes and Sanremo, there are media reports which suggest localised pluvial flooding (http://www.nicematin.com/diaporama/intemperies-les-photos-des-internautes.1971025.html?idx=9#top-diapo) but not fluvial flooding. There is one observed flooding location at Bordighera in Liguria but this related to the previous day, hence the warning points were one day too late. In upper Switzerland the precipitation fell as snow and no warning points or flooding were reported. Further downstream the ERIC produce warnings in Lombardy near the Swiss border, something which is corroborated by a report in the European Severe Weather Dataset at Val Masino. Alerts are generated in ERIC along the Adige and Po rivers, however no flooding was observed along these reaches. These points were forecast to occur in the latter half of the day when the precipitation was forecast to spread into this area. The most significant observed flooding in Italy was at Carrara on the coast near La Spezia. ERIC produced two warning points 68 km to the north-west of this location

Fewer warning points are produced for the 6/11/2014 by ERIC those generated are clustered along the Adige river in Trentino. These events were forecast to occur in the early hours of 6th November 2014, this may have resulted from an increase in the forecasted precipitation in the COSMO run on 2014-11-05-00. However, no flooding was observed in these areas. There is evidence that the COSMO-LEPS NWP product used in ERIC over predicted the precipitation in this region, so it is unlikely that the cause of these erroneous warnings is due to the precipitation falling as snow.

5.4 Algorithms for flash flood nowcasting

5.4.1 Uncertainty for flash flood nowcasting

Flash flood forecasting based on the FF-EWS module (Corral et al., 2009; Alfieri et al., 2011; 2017) characterizes the flood hazard based on the catchment-aggregated rainfall (i.e. the rainfall integrated in the catchment upstream of each point of the drainage network and over its characteristic concentration time). The module uses the high-resolution precipitation inputs generated from radar observations and nowcasts.

Given its simplicity, (the FF-EWS assesses the flash flood hazard by means analyzing the main forcing i.e. the catchment-aggregated rainfall), the uncertainty in the FF-EWS forecasts is characterized using high-resolution probabilistic rainfall inputs.

The probabilistic version of the FF-EWS module uses the rainfall forecasts obtained with the ensemble nowcasting technique SBMcast (Berenguer et al., 2011).

SBMcast runs every time new radar Quantitative Precipitation Estimates are available. It assumes a statistical model for the precipitation field (the String of Beads model, Pegram and Clothier 2001) to evolve from the initial conditions while

preserving the spatial and temporal structures as well as the motion of the precipitation field. The result is an ensemble of possible rainfall nowcasting scenarios (for up to 3 hours) coherent with the radar observations.

In this way, the results of the probabilistic version of the FF-EWS module are the forecasted flash flood return period at each point of the drainage network together with the probability to exceed different return periods, estimated from the results obtained by the different members.

5.4.1.1 Probabilistic flash flood nowcasting in Catalonia

In the case of the Catalonia Pilot Site, the FF-EWS module is integrated in the operations of the Catalan Water Agency and uses the available hydrometeorological and the radar QPEs produced by the Catalan Weather Service.

The example below corresponds to the event of 02 November 2008 during which a convective system moved toward the interior of Tarragona resulting in high rainfall intensities and some flooding in several streams. The FF-EWS module was able to identify significant hazard level in some coastal torrents (Figure 11), where the most damaging flash floods occurred. Figure 11 shows the evolution of the hazard map forecasted for 06:00 UTC. The first signal of possible hazard in this area was detected 1.5 hours ahead (i.e. with the forecast generated at 04:30). Thirty minutes later, the deterministic hazard map forecasted with a leadtime of 1 hour is almost identical to what was finally diagnosed from radar observations (Figure 11a).

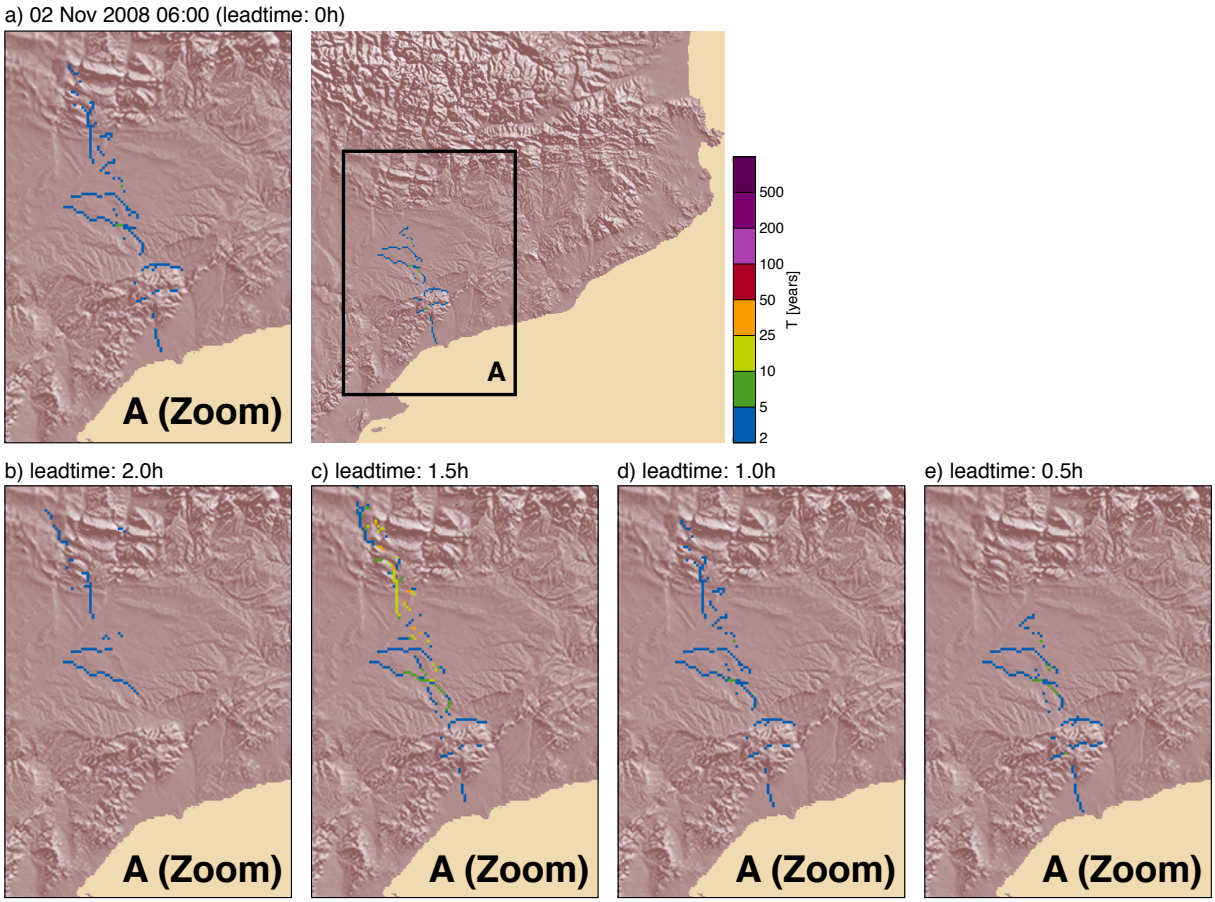


Figure 11: (a) Flash flood hazard estimated on 02 November 2008 at 06:00 UTC. (b-e) Evolution of the hazard assessment product for leadtimes of 2.0, 1.5, 1.0 and 0.5 hours.

The second example (see Figure 12) corresponds to the case of 12-14 September 2006, a more generalized event that resulted in rainfall accumulations exceeding 200 mm in 24 hours in several places of the domain, producing flash floods in ephemeral torrents near the coast and in some sub-basins of the main rivers.

Figure 12 shows the variety of hazard assessment products generated by the system every time a new radar QPE map is available: besides of the hazard assessment based on radar observations (Fig. 12a), the system forecasts the expected evolution of the hazard level based on radar QPF. Additionally, it estimates the uncertainty of the forecasted hazard level in terms of probability to exceed a given return period (Figs. 12c and 12d show, respectively, the probability to exceed a return period of 2 and 5 years). This product describes how the uncertainty in rainfall nowcasts affects the forecasted hazard.

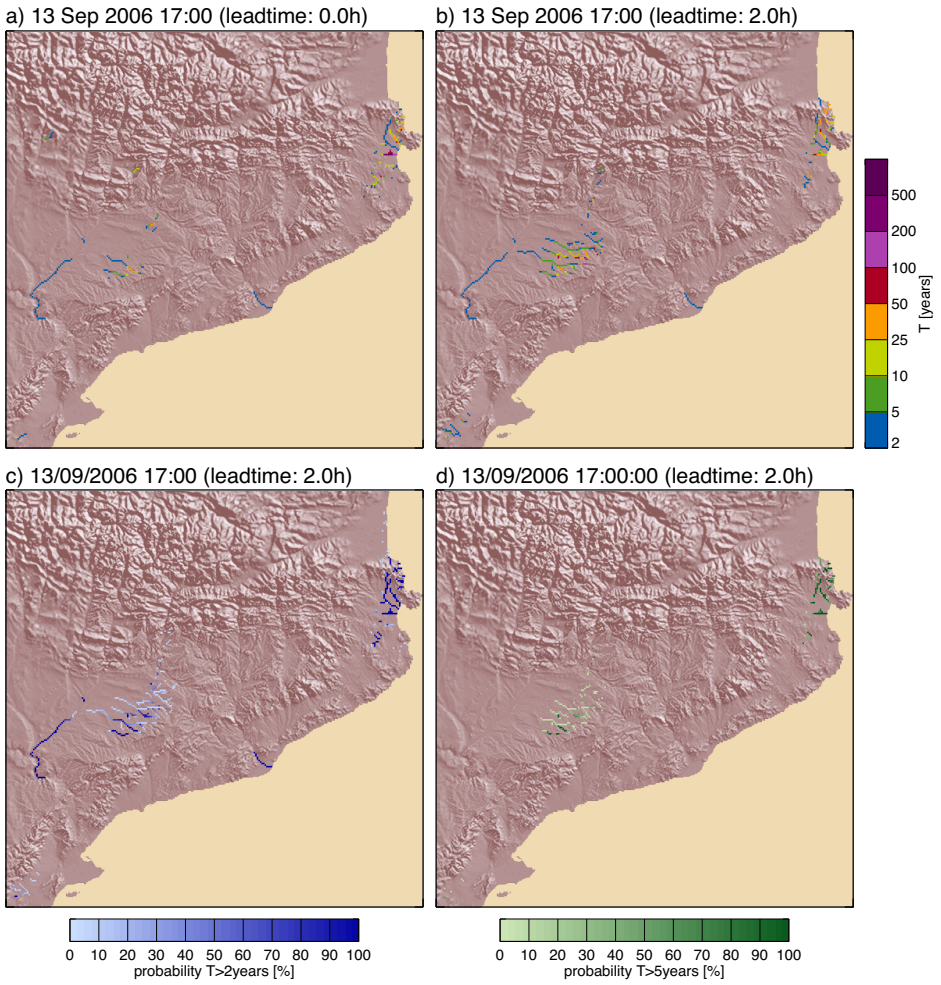


Figure 12: Hazard assessment for 13 September 2006 at 17:00UTC: (a) obtained radar rainfall observations; (b) obtained for a leadtime of 2 hours. (c) and (d): Respectively, probability of exceeding

a return period of 2 and 5 years, as estimated from probabilistic rainfall forecasts with a leadtime of 2 hours.

5.4.1.2 Robustness for flash flood nowcasting

Robustness of the very-short term flash flood forecasts relies on the ability to estimate the return period of the observed and forecasted catchment-aggregated precipitation. In a changing climate, this would then be the most affected element of the FF-EWS and would require re-estimating the maps of catchment-aggregated rainfall for different return periods. It is likely that extreme precipitation will increase, especially during winter, spring and autumn due to non-convective, large-scale precipitation episodes. This fact would result in an increase of rain-on-snow events leading to floods. This should be in future consider to improve robustness.

5.5 Algorithms for Storm Surge forecasting

5.5.1 Uncertainties of the European Storm Surge model

The verification of the European storm surge model is carried out comparing the computed storm surge against water level time series available from the JRC database of tidal gauges (Figure 13). The tidal data-set consists of more than 200 tide gauges, covering the whole computational domain. The temporal resolution of sea level measurements varies from minutes to 1 hour while the temporal coverage varies from one year up to 35 years.

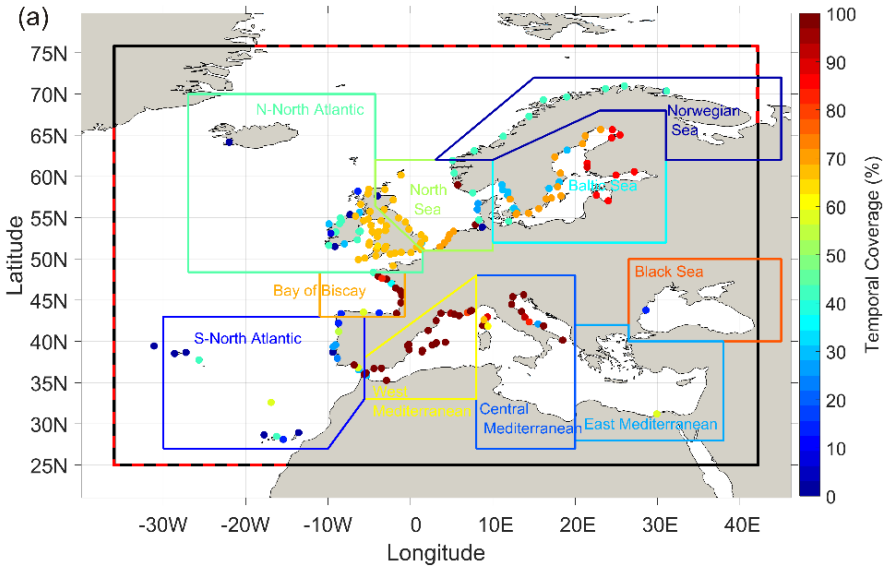


Figure 13. (a) Map of Europe showing the Schism computational domain (black line), open boundaries (black-red dashed line), 10 different coastal regions defined for the analysis of the model results (colour continuous lines), and the location of tidal gauge stations (the colour scale shows the temporal coverage of the tidal gauges for the validation period from 2010 to 2016).

For a long-period, multiannual verification, that will include a large number of extreme events, the Era-Interim (EI) atmospheric forcing (mean sea level pressure and wind velocity) developed by the ECMWF is used. This dataset extends from 1979 to 2016. Since the operational storm surge forecast uses the high resolution atmospheric forecast (HR) provided by ECMWF, a comparative analysis of the model performance using the HR and EI will be carried out for a period spanning from 2010 to 2016.

The assessment of the algorithm performance was evaluated in terms of the root mean square error (RMSE) (eq.1), relative root mean squared error (%RMSE) (eq. 2) and Pearson correlation coefficient (r) (eq. 3).

$$RMSE = \sqrt{\frac{\Sigma_k^n (\eta_o^k - \eta_p^k)^2}{n}} \qquad (1)$$

$$\%RMSE = \frac{\sqrt{\frac{\Sigma_k^n (\eta_o^k - \eta_p^i)^2}{n}}}{\max(\eta_o)} \cdot 100 \qquad (2)$$

$$r = \frac{\Sigma_i^n (\eta_o^k - \overline{\eta_o^k})}{\sqrt{\Sigma_i^n \left(\eta_o^k - \overline{\eta_o^k}\right)^2} \sqrt{\Sigma_i^n \left(\eta_p^k - \overline{\eta_p^k}\right)^2}} \cdot 100 \quad (3)$$

Where n is a number of measurements in the time series at a given location, η_o is the observed storm surge, η_p is the predicted storm surge.

Figure 14 shows the spatial variation of the accuracy of the European Storm surge model prediction, according to the error parameters defined above. A statistic summary organised by oceanographic regions can be found in Table 5.

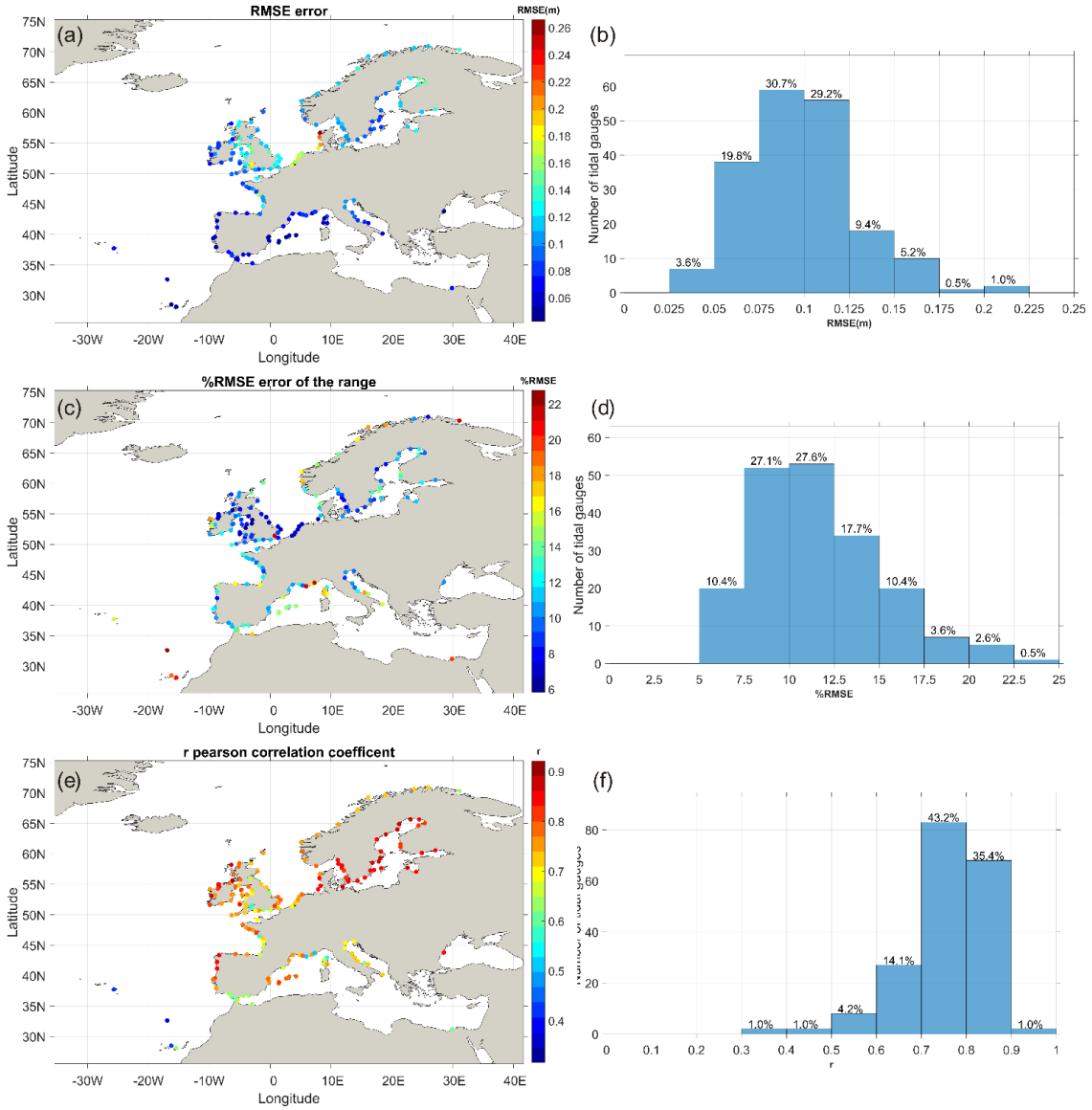


Figure 14. Model validation performance for the storm surge in all the tidal gauge stations considered. (a) Map plot of RMSE; (b) Histogram of the RMSE; (c) Map plot of %RMSE; (d) Histogram of the %RMSE; (e) Map plot of the r correlation coefficient; (f) Histogram of the r correlation coefficient.

Table 5 Statistics of model performance along the coastline of the 10 defined European regions

	RMSE (m)				%RMSE (%)				r			
Region	Mean	Std	Min	Max	Mean	Std	Min	Max	Mean	Std	Min	Max
Black Sea	0.04	-	0.04	0.04	10.59	-	10.59	10.59	0.83	-	0.83	0.83
East Med.	0.07	-	0.07	0.07	19.81	-	19.81	19.81	0.59	-	0.59	0.59
Central Med.	0.08	0.02	0.05	0.10	13.26	3.36	8.59	18.39	0.69	0.06	0.54	0.76
West Med.	0.06	0.01	0.05	0.08	14.55	3.18	10.03	21.36	0.71	0.08	0.52	0.81
S-North Atlantic	0.06	0.01	0.04	0.08	13.68	4.50	8.06	22.75	0.64	0.17	0.35	0.85
Bay of Biscay	0.09	0.03	0.06	0.16	12.22	2.34	8.51	16.40	0.74	0.07	0.58	0.84
N-North Atlantic	0.11	0.03	0.07	0.19	9.85	3.18	5.83	21.85	0.76	0.07	0.58	0.92
North Sea	0.14	0.05	0.07	0.27	10.11	3.45	5.87	17.03	0.76	0.05	0.63	0.85
Baltic Sea	0.10	0.02	0.08	0.16	10.42	2.08	7.43	15.12	0.84	0.08	0.32	0.90
Norwegian Sea	0.12	0.01	0.11	0.13	15.15	4.55	7.58	21.60	0.72	0.04	0.65	0.77

5.5.2 Uncertainties related to extreme surge prediction

The verification of the algorithm in the case of the extreme surges prediction was undertaken only for the tide gauge stations that contain time series longer than 10 years. In those stations, an extreme value analysis was carried out to compare measured and modelled extreme storm surge levels. The peak-over threshold (POT) approach was followed, extremes were selected taking a 3-day time windows to consider independent events, and the threshold corresponding to the 95.5th percentile. The selected exceedance events per year were modelled according to the Generalized Pareto Distribution (GPD). In addition, the BIAS was calculated for each return period levels (Tr= 5, 10, 20, 50, 100).

As an example, Figure 15 shows the value of BIAS calculated for storm surge level corresponding to the 5 (top panel) and 10 (bottom panel) years return period.

Figure 16 shows the GPD plot for the modelled and observed extreme storm surges at selected stations. In general, a good performance was observed. However the model underestimates extreme surge events at some locations.

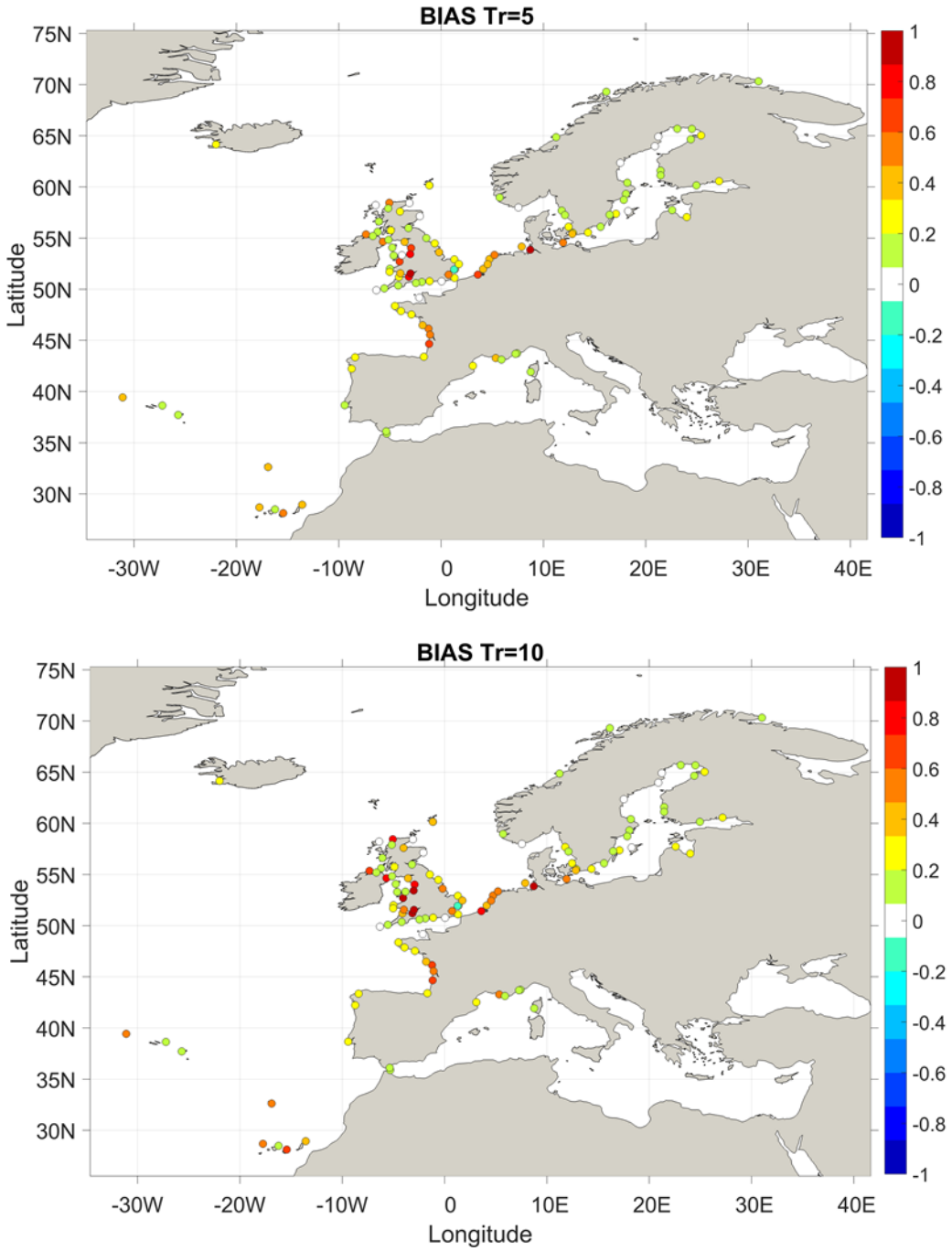


Figure 15 The BIAS (m) calculated by subtracting the modelled extreme sea levels from the observed extreme sea levels with a return period of 5 years (upper panel) and with a return period of 10 year (bottom panel).

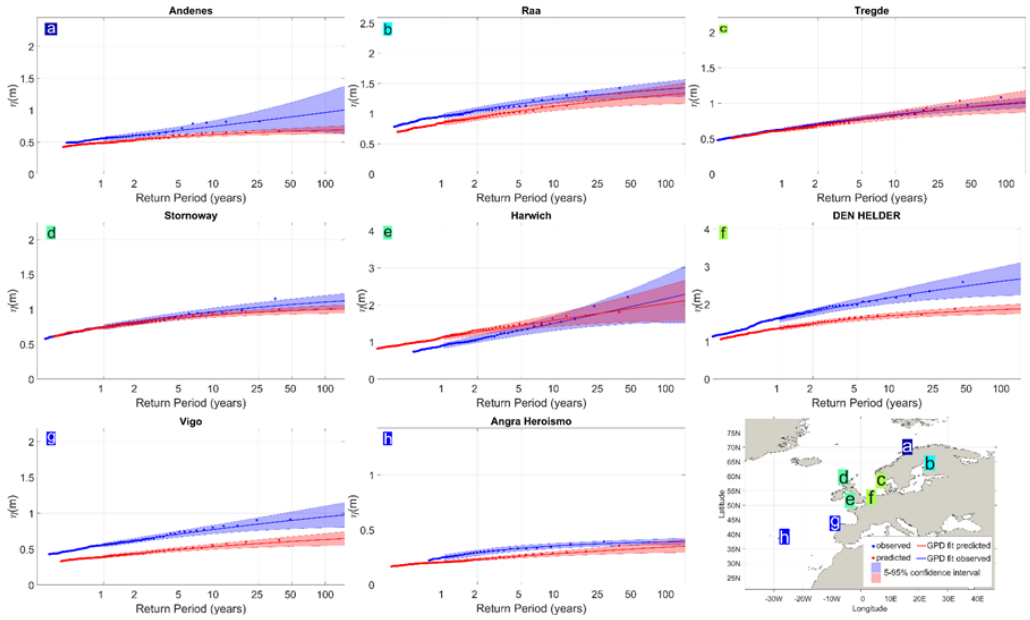


Figure 16. General Pareto Distribution plot for the modelled and observed extreme storm surges at selected stations.

5.5.3 Astronomic tide verification

In the case of pure tidal signal, the model performance was evaluated in terms of absolute error (ε_A) (eq.4), relative error (δ_A) (eq.5) and absolute error for the tidal phase (ε_P) (eq.6). The main tidal constituents were analysed.

$$\varepsilon_a^i = a_o^i - a_p^i \qquad (4)$$

$$\delta_a^i = \frac{a_o^i - a_p^i}{a_p^i} \qquad (5)$$

$$\varepsilon_p^i = p_o^i - p_p^i \qquad (6)$$

Where *a* is the tidal amplitude and p is tidal phase, subscript index o is referred to the observed data, and subscript index p is referred to the predicted data.

Figure 17 shows the verification of the main semidiurnal tidal constituent M2 and S2. The largest relative errors were observed in micro-tidal areas where the tidal signal contribution to the total water level is minimal

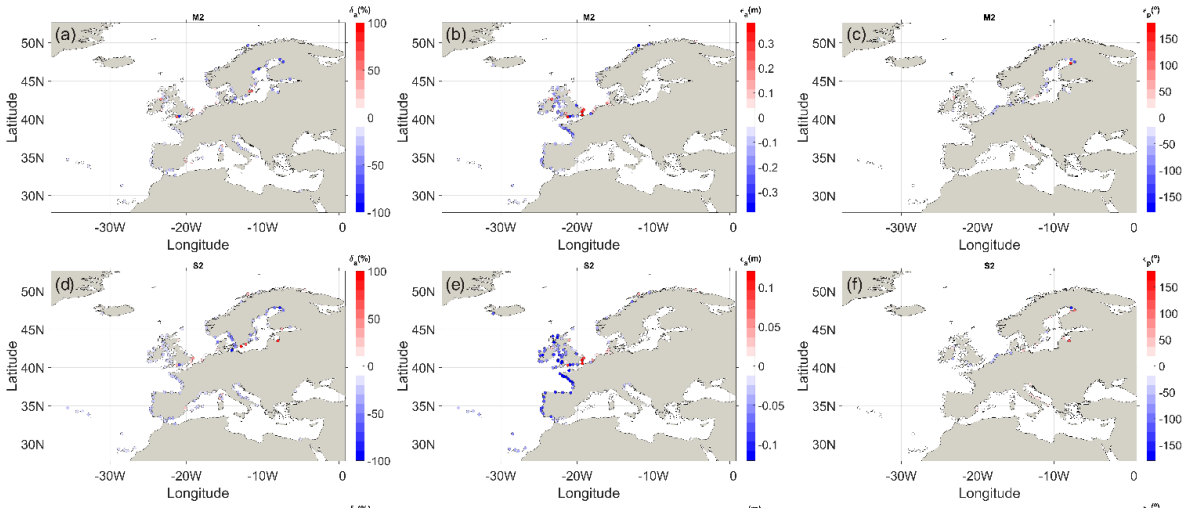


Figure 17. Spatial variation of the tidal validation result for the principal semidiurnal tidal constituent (M2, S2). Relative error in amplitude (left panels). Absolute error in amplitude (centre panels). Absolute error in phase (right panels).

Figure 18 shows the histogram of the absolute and relative error of tidal amplitude and absolute error in tidal phase. Overall, acceptable results were achieved in most of the tidal stations. However, the quality of the tidal reduction will be improved in the future thanks to an increase in mesh resolution in areas with complex bathymetry.

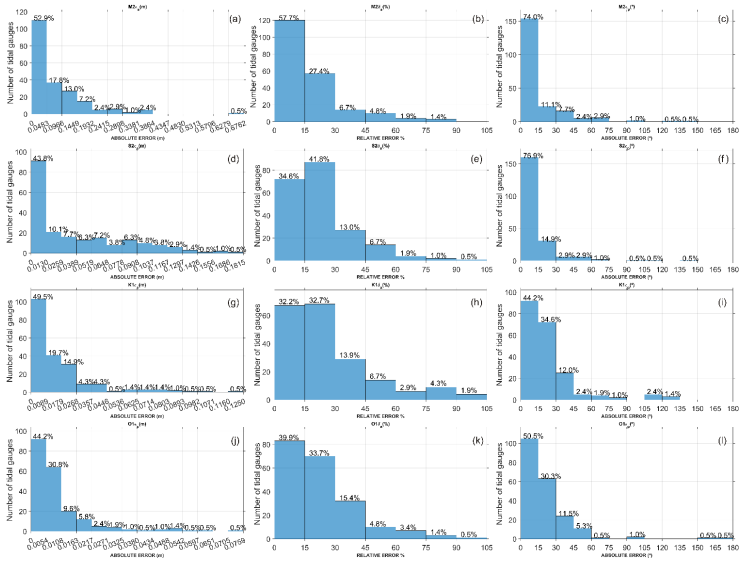


Figure 18. Histograms of the of the model error for the principal semidiurnal tidal constituent (M2, S2) in all the tidal gauge station considered. Absolute error in amplitude-left panels. Relative error in amplitude-center panels. Absolute error in phase-right panels.

5.5.4 Robustness and alert analysis

The Storm surge hazard level is calculated based on the exceedance of the forecasted extreme surge level corresponding to different return periods of the storm surge (5, 10, 20, 50 and 100 years). The system alerts when the forecast storm surge level is greater than the surge level corresponding to a 10-year return period. An analysis of the alerts counted as hits or false alarms will be done comparing the

algorithm output with the tide gauge measurements. The hit rate (HR) is calculated according to eq. 7.

$$HR = \frac{\sum hits}{\sum hits+misses} \qquad (7)$$

Where: “hits” accounts for the correct predicted alert verified in the observation data and “misses” accounts for a no alert that was instead verified as present in the observation data.

Additionally, another index that can be used to verify the storm surge forecast is the success ratio (SR) eq. 8:

$$SR = \frac{\sum hits}{\sum hits+falert} \cdot 100 \qquad (8)$$

Where “falert” accounts for the false alarms.
Figure 19 illustrates an example of HR calculated in the area of the Baltic Sea.

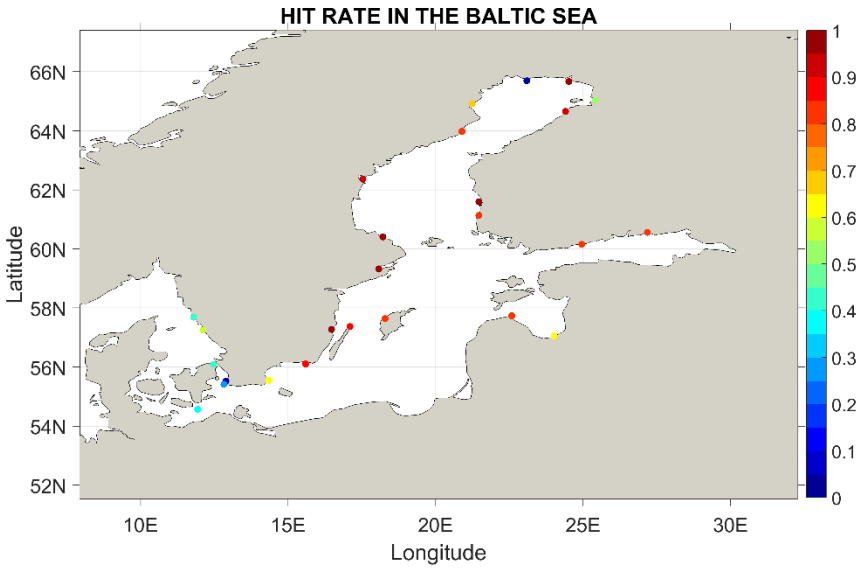


Figure 19 Example of Hit Rate calculated in the Baltic Sea.

5.6 Algorithms for Drought Forecasting

5.6.1 Uncertainty for Drought Forecasting

Drought forecasting stands between medium-range forecasting, which is strongly related to initial conditions, and the seasonal time-scale, mainly driven by oceanic variability and large-scale climate features such as the El-Niño. Such forecasts are highly uncertain due to the chaotic nature of the atmosphere (Vitart, 2014). Only a few studies are available that uncertainty in seasonal drought forecasting. Some encouraging initiatives are ongoing, that is, at the national scale (Prudhomme, 2015; Meißner et al., 2015; Ionita et al., 2008; 2015), and at the pan-European scale (Lavaysse et al., 2015; Arnal et al., 2017). The ANYWHERE project will contribute to the knowledge of uncertainty in drought forecasting by testing of the seasonal drought forecasts with the observed impacts in the ANYWHERE Pilot Sites. Below, we will show some preliminary results for the Pilot Site Catalonia, which we obtained in co-operation with the Catalonian Water Agency (ACA).

5.6.1.1 The case of seasonal drought forecasting for Catalonia

Three resources are used for water supply in Catalonia, i.e. (i) surface water reservoirs, (ii) aquifers, and (iii) desalinisation of sea water. ACA prefers to use the water stored in the reservoirs, because of the high costs of the desalinisation and the salt-water intrusion in the coastal aquifers. Twice a year, the Catalonian Water Agency (ACA) forecasts the water volume for each reservoir for the upcoming 6 months. The volume is based upon the forecasted probabilistic river inflow using historic flow data and estimated water demands. This information is shared and discussed with stakeholders (Comissió) in spring and autumn of each year. Some recent examples are presented in Figure 20. The colours indicate warning levels (light yellow: alert, dark yellow: exceptional, and brown: emergency). Each level triggers different measures. The ACA forecast for the Ter-Llobregat reservoir system that was issued in April 2017 is given in Figure 20 (left). The forecasted volumes are based on the initial reservoir volume and river flow that has been observed in the past. These graphs are named a “spaghetti diagram”.

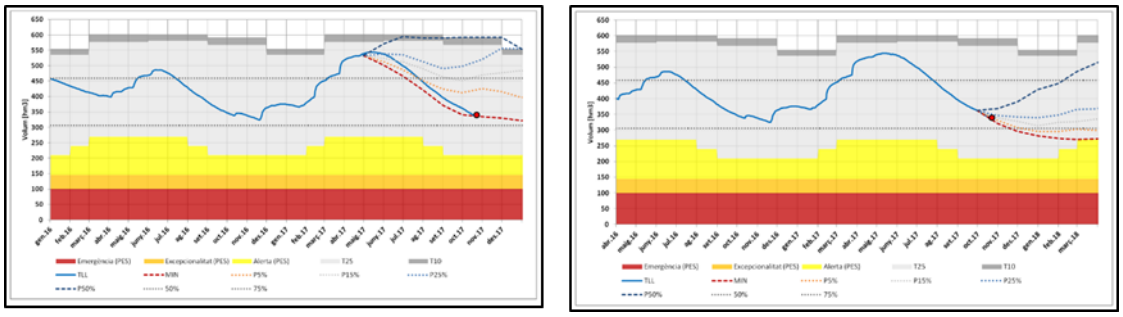


Figure 20: Observed (blue line) and forecasted (dashed lines) storage in the Ter Llobregat reservoir system (Catalonia). Left: multi-monthly forecast on April 2017 and realisation, Right: probabilistic forecast for period October 2017-March 2018. The colours indicate the different warning levels (Source: Catalonian Water Agency, ACA).

The forecasted reservoir volume is probabilistic of nature, i.e. 5, 15, 25 and 50th percentiles are given, as well the minimum storage. Spread in the forecasted reservoir storage by the end of October 2017 is large. The forecasted median storage is about 600 Hm^3, whereas the 5% storage is somewhat higher than 400 Hm^3. The left panel of Figure 20 also provides the observed storage and shows in this case that the storage by the end of October was close to minimum storage. The ACA forecast for the upcoming 6 months until March 2018 is illustrated in Figure 20 (right panel). The reservoir evolution in the first 20 days in October still follows the minimum scenario. The spaghetti diagram clearly shows that the situation at the beginning of the summer of 2018 would already become critical (close to alert level) if the 5th or 15th scenarios would happen. The median scenario (50th scenario) would offer relief. More concrete information on possible evolution of the reservoir volumes would be extremely valuable for the Catalonian Water Agency.

5.6.1.2 ANYWHERE seasonal drought forecasting for Catalonia

ANYWHERE provides through the MH-EWS seasonal drought forecasting products for Catalonia at 5 km spatial resolution.

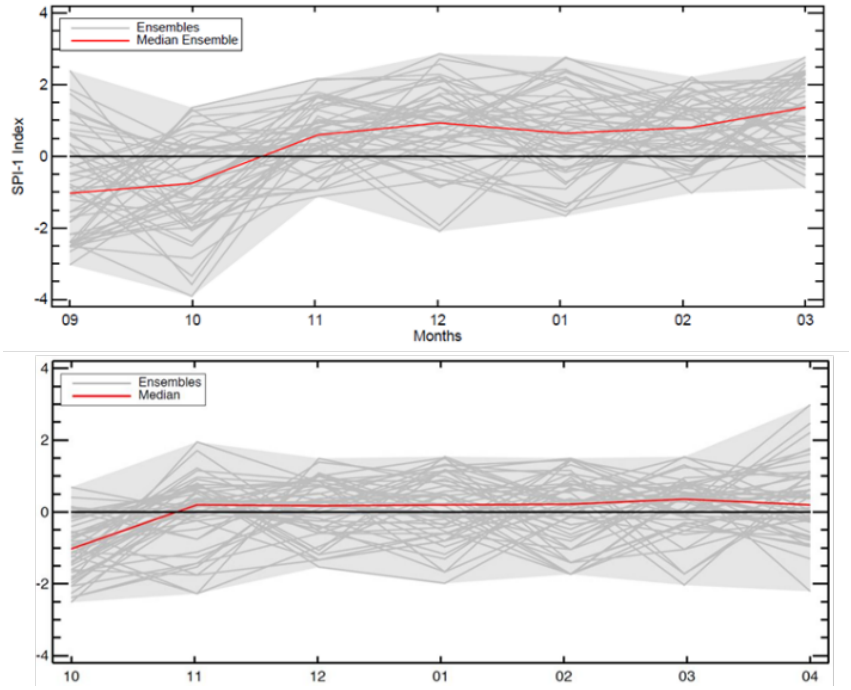


Figure 21: Forecasted probability of drought in precipitation (meteorological drought) using SPI-1 (median and 51 ensembles) in Catalonia (7 months ahead. Upper: September 2017 – March 2018) (forecast September 2017), and Lower: October 2017 – April 2018) (forecast October 2017)

The forecasted drought in precipitation, through the Standardized Precipitation Index accumulating precipitation over 1 month (SPI-1) reflecting meteorological drought is shown in Figure 21. The forecasted median SPI-1 for October is about -1 meaning that the forecasted precipitation is 1 standard deviation lower than the median of October precipitation (Figure 21, upper). This implies dryer conditions lower than normally in Catalonia, which is confirmed by the decrease in reservoir storage (Figure 4, right). The forecasted spread for the October SPI-1, however, is rather

large ranging from about -2.5 to +0.5 (substantially lower precipitation than normal to slightly wetter than normal). The September seasonal forecast reveals a slightly upward trend (Figure 5, upper), that is indicating the development of wetter conditions from November onwards. However, there are still some ensemble members that forecast a below-median precipitation for the first months in 2018. The October seasonal forecast of SPI-1 (Figure 21, lower) more or less confirms the September seasonal forecast. However, the upward trend towards wetter conditions from November onwards is not so clear anymore when the ensemble spread is taken into account.

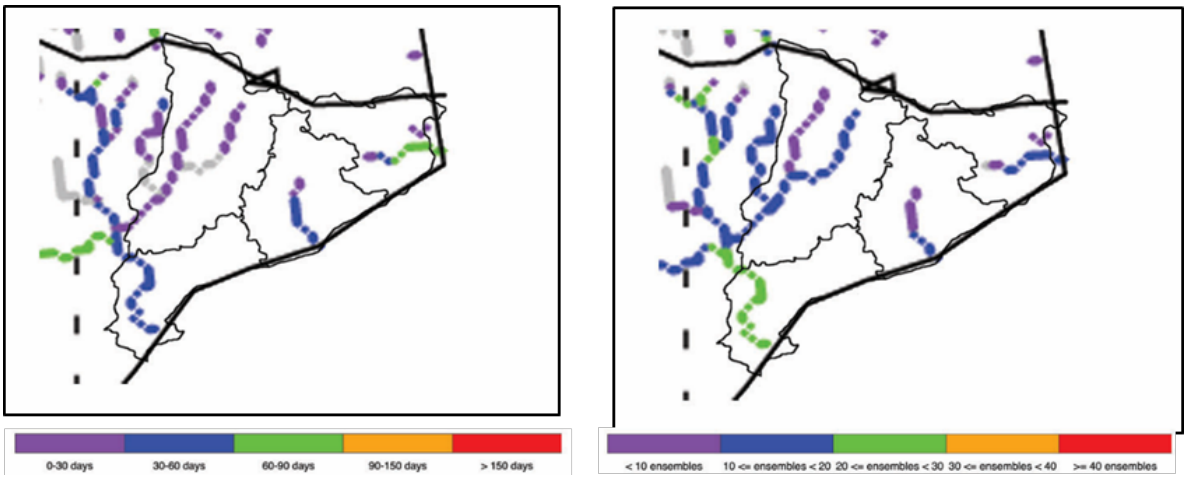


Figure 22: Forecasted drought in river discharge in Catalonia (hydrological drought) using the variable threshold approach. Left: 50% probability of the total drought duration in days (7 months, September 2017 - March 2018), and Right: number of ensemble members in drought by the end of November 2017 (forecast September 2017).

The forecasted drought in river discharge (hydrological drought) in Catalonia is more relevant for the inflow in the reservoirs than the drought in precipitation (meteorological drought), as presented in Figure 22. In the next phase of ANYWHERE forecasted seasonal hydrological drought will be considered. Forecast skills will be evaluated.

5.6.2 Robustness for Drought Forecasting

Robustness of seasonal drought forecasts has hardly been investigated. Identification of drought from the forecasted time series of hydrometeorological variables using standardized drought indices depends on the pre-selected probability distribution. McKee et al. (1993) provide information for the Standardized Precipitation Index (SPI), whereas Bloomfield et al. (2013) and Shukla and Wood (2008) give this for the Standardised Groundwater Index (SGI) and the Standardized Runoff Index (SRI). Selection of the threshold to identify drought using the Threshold Level Approach affects the ANYWHERE drought products. For instance, selection of a more extreme threshold (i.e. 90th instead of 80th percentile) usually leads to shorter and less severe drought events (e.g. Tallaksen and Van Lanen, 2004; Heudorfer and Stahl, 2017).

Drought forecasting under a future climate (droughts are expected to increase across Europe but most severely over its southern parts; see section 4.2.1) is anticipated to be rather robust, although it is relevant whether people adapt or not to droughts is essential. Wanders et al. (2015) illustrate that a gradual adaption to a changing hydrological regime has a significant influence on the duration and severity of drought events under a future climate.

5.7 Algorithms for weather-induced heatwaves and related health impacts

5.7.1 Uncertainty in weather-induced heatwaves and related health impacts

The uncertainty of UTCI forecast to ECMWF inputs variables – 10 m wind speed, 2 m relative humidity, 2 m temperature, and solar radiation – and their variability was investigated by Pappenberger et al. (2015). Scatterplots were used for the purpose, showing the relationship between computed UTCI values and each meteorological input variable for the entire reanalysis period (1979-2009) and all grid points (global scale) (Figure 23). The main outcomes were that the UTCI is sensitive to all input parameters with (a) some linear dependencies on air temperature, (b) a distinct lower boundary for wind (> 17 m/s) justified by the clothing insulation and vapour resistance caused by body movements and the wind itself (Havenith, et al. 2012), and (c) a lower boundary influenced by the solar elevation angle, solar radiation and thermal radiation.

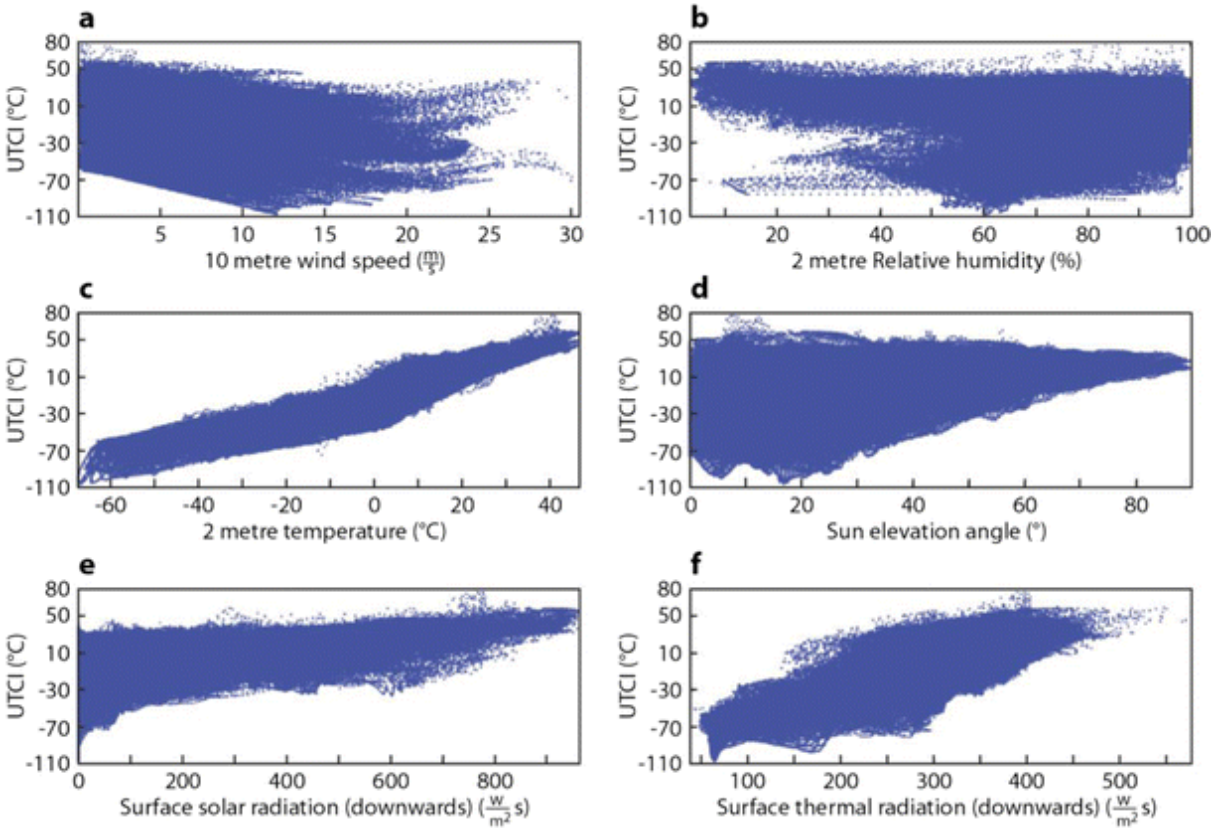


Figure 23 Scatterplots of meteorological inputs against UTCI to illustrate associated dependencies (Pappenberger, et al. 2015).

5.7.2 Robustness in weather-induced heatwaves and related health impacts

Similarity between forecast and observed UTCI was also analysed by Pappenberger et al. (2015). UTCI forecasts were calculated every day with a lead time of 10 days from 1 January 2009 to 31 December 2012 using both the ECMWF HRES and ENS inputs. The deterministic high-resolution, control and ensemble mean forecasts of the UTCI were then compared with observation using the Anomaly Correlation Coefficient (ACC). Pappenberger et al. found that the maximum lead time for which ACC stays above 60% is: 4-6 days in the Mediterranean Basin (30°N–48°N, 10°W–40°E); and 2-4 days for deterministic high resolution and control forecast, and 4-6 days for the ensemble mean forecast in Northern Europe (48°N–75°N, 10°W–40°E) (Figure 24). Both geographical context are expected to suffer a major increase on average temperature and heatwaves (see section 4.2.1)

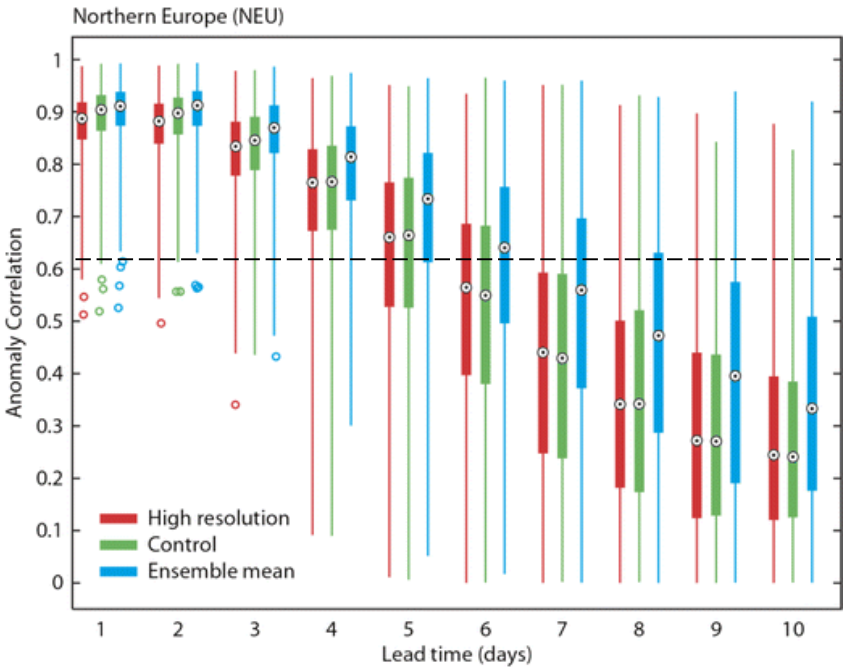


Figure 24 Anomaly correlation for the Northern European area and the three different UTCI forecasts. The circle illustrates the mean whilst the box indicates the 25th and 75th percentile. The whiskers of the box plot extend to the 95th and 5th percentile. Coloured circles indicate outliers. (Pappenberger, et al. 2015).

The skill of UTCI forecasts in the prediction of strong heat stress (>32°C) was assessed via the Brier Skill Score (BSS). The lead time at which the BBS drops below zero, i.e. there is no skill compared to climatology, is highest for ensemble predictions rather than the control and the high-resolution forecasts (Pappenberger, et al. 2015).

5.7.2.1 An example: the Russian heatwave of summer 2010

In summer 2010, a blocking anticyclone over western Russia drove warm air from Africa rising temperatures to unprecedented levels, e.g. 38.2ºC in Moscow (Ghelli, et al. 2010). The 2010 Russian heatwave caused a death toll of 55,000 (Barriopedro, et al. 2011). From 10 days before the event, the UTCI ensemble forecast shows a

significant probability for strong heat in Moscow between 25th and 28th July (Figure 25). For example, the ensemble forecast issued on 16 of July is green for the 23th July meaning that up to 25% of the ensemble members had a UTCI exceeding 32°C, and yellow for the following days, showing an increase probability of up to 50%. The signal for the event becomes stronger in the later forecasts. The high resolution and control also indicate a possible event, although at the longer range, the timing of this is not consistent from day to day (i.e. the signal flip-flops, or changes). These results show the potential of UTCI forecast as an early warning indicator of a heatwave (Pappenberger, et al. 2015).

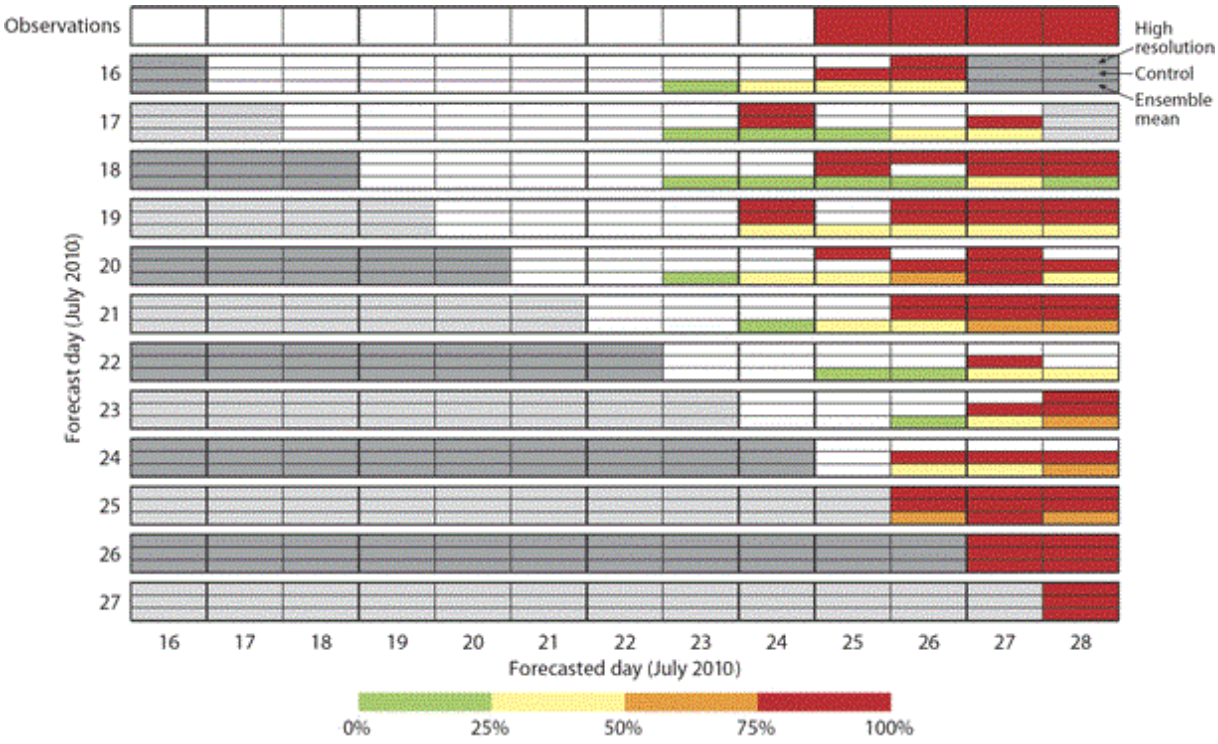


Figure 25 UTCI forecasts for Moscow during the Russian heatwave of summer 2010. The top line represents the observations. Red indicates days where the UTCI exceeds 32 °C. All other lines show forecasts issued on particular days. For example, the second line shows a forecast issued on the 16th July. Each forecast is sub-divided into three sub-boxes. The top sub-box shows the high resolution (red when the forecast UTCI exceeds 32 °C). The middle sub-box shows the control (red when the forecast UTCI exceeds 32 °C). The bottom sub-box shows the ensemble; the percentage of ensemble members with UTCI exceeding 32 °C is colour coded (Pappenberger, et al. 2015).

5.8 Algorithms focused on wildfire danger modelling

Forecasting wildfires is a complex task as fires depend on a stochastic (unpredictable) component, namely ignition. The trigger can be natural in origin like lightning and self-combustion. But it can also be due to human behaviour as intentional act of a person or unintentional act of negligence. Forecasting wildfire risk means quantifying a 'potential' risk, not a real one, therefore this type of hazard is intrinsically dominated by false alarms. It is when a fire is ignited that modelling risks becomes more valuable.

In order to understand what can be predicted, we need to keep in mind that flames tend to rage out of control if certain soil and atmospheric conditions are met. Therefore, fire prediction systems (like the Global ECMWF Fire Forecasting, GEFF) are designed to highlight these favorable weather conditions which can allow sustained fire activity. On these premises is based one of the most widespread fire danger rating system, the Canadian forest service's Fire Weather Index (FWI) (Van Wagner et al. 1987, 1974) which is selected here to showcase our first attempts to quantify GEFF's outputs uncertainty and robustness.

5.8.1 Uncertainties on wildfire danger modelling

The FWI is a measure of fire potential and is expressed as a numeric rating. Rating rises as fire weather becomes more severe. However, the same FWI values can correspond to different danger levels depending on the ecosystem. To become a meaningful tool in fire management, the FWI requires the definition of danger levels that should be site specific. Vitolo et al. (2017 a, b) describe how to calculate these danger levels for Europe, at various scales (from pan-European to country, region and province) and propose the use of a post-processing tool called 'Caliver' to make the analysis easily reproducible. Table 6 summarizes FWI danger levels for selected areas in Europe. The first row shows the current pan-European levels as defined by EFFIS. The remaining rows list the danger levels as calculated by Caliver, the values in bold are used for validation in the next phase.

Once danger levels are defined, we compare the risk level with actual burned areas and quantify the number of hits and misses (as, mentioned before false alarms are irrelevant in the wildfire context) and the resulting probability of detection and Area Under the (ROC) Curve.

According to EFFIS, the high danger threshold is 21.3 which corresponds to a probability of detection of 47% and an AUC score of 0.718. Repeating the same exercise using caliver's threshold for Europe (10), the probability of detection increases to 65% and the AUC score to 0.781.

Figure 26 shows in black the ROC curve corresponding to EFFIS' threshold and in red the curve corresponding to Caliver's threshold.

Table 6 - FWI danger levels for selected areas. Values in bold are used for validation.

Area of interest	Low	Moderate	High	Very high	Extreme
Europe (EFFIS - current standard)	5.2	11.2	**21.3**	38	50
Europe	2	5	**10**	19	33
United Kingdom	1	3	**7**	12	19
Spain	2	7	**15**	30	55
Italy	2	6	**12**	23	42
Calabria Region (IT)	2	6	**12**	24	42
Sicily (IT)	2	6	**14**	27	49
Liguria Region (IT)	2	4	**9**	16	28
Province of Genoa (Liguria Region, IT)	2	4	**9**	17	29

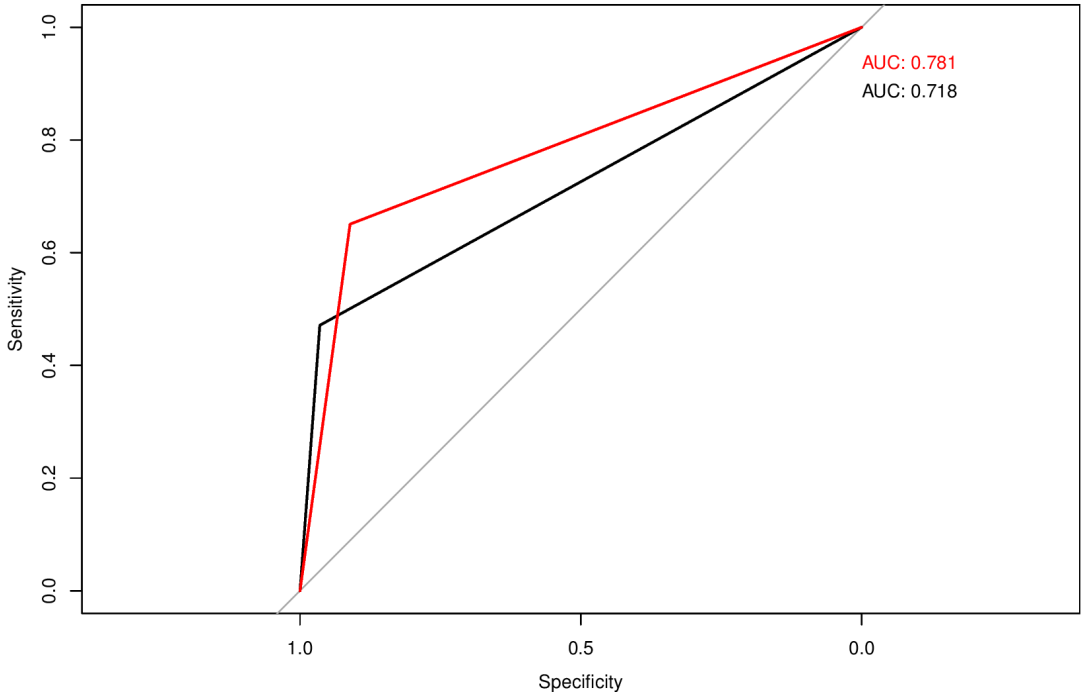


Figure 26 - ROC curves and AUC scores derived from the validation of EFFIS standard thresholds (black) and Caliver (red) newly calibrated thresholds.

If country-specific danger levels are considered, instead, the high danger threshold fluctuates between 5 and 15. The procedure above can be iterated for the above selected areas in Europe so that hits and misses can be calculated country-by-country. Table 7 summarizes the output of this validation using EFFIS and Caliver's danger levels. Compared to EFFIS, Caliver's methodology returns a systematically higher number of hits and lower number of misses, both at the European and country level scales (for United Kingdom, Spain and Italy). Comparing European and country-specific danger levels, the latter represents a further improvement only in the northern countries (i.e. UK) whereas it seems generally better to consider European levels for Mediterranean countries.

Table 7 - Comparison of hits and misses using various danger levels: EFFIS, Caliver's levels over Europe and Caliver's country-specific levels. Caliver's methodology systematically returns higher number of hits and lower number of misses. The last column shows hits and misses considering Europe as the sum of its parts.

		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

More in general, Figure 27 shows the probability of detection at the global level for large regions at day 6 forecast. POD seem to perform rather well in Africa, North and South Americas, Europe, Asia and Australia. The lowest POD is estimated in central America and Tropical Asia, this result requires further investigations.

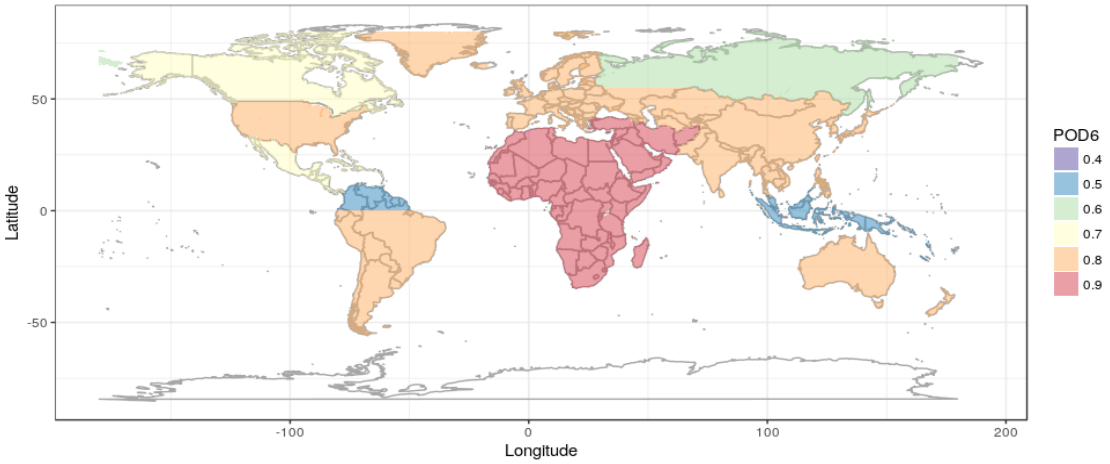


Figure 27 - Global area averaged Probability of Detection (POD) for large regions.

5.8.2 Robustness on wildfire danger modelling

As the last step, we attempt to build operational confidence by using the probabilistic forecast and visually assess the spatial variability of the distribution of danger and how this diverges from the deterministic estimate. This spatial investigation is still work in progress.

Currently we do not test GEFF on future climate scenarios, as the estimation of fire risk is more valuable at short to medium range and once the ignition as already taken place.

5.9 Algorithm for Storms forecasting

In order to enable risk estimation of the natural hazards associated to various types of storms many of the NWP models can presently provide probabilistic ensemble predictions or such can be obtained from deterministic model output by time-space-model lagging. Exceedance probabilities of present threshold values of natural hazard severity classes are mandatory for estimating socio-economic impacts and human fatalities. As a useful "side effect" uncertainty of NWP model forecasts can be estimated from the width of the probability distribution of the ensemble prediction. A well-established set of quantitative metrics has been developed in the NWP community for the verification of the skill of both deterministic and probabilistic predictions (e.g. Magnusson et al., 2014). On the other hand, the impact algorithms of natural hazards and their verification methods are less developed than those with the NWP. The robustness of NWP model algorithms in a changing climate, with many uncertainties in it, is not so poor as one might think. During the past decades there has been a continuous development among the NWP models that has led to their constantly improved skills. It can be expected that during the coming decades model parameters (e.g. CO_2 concentration), changing environmental circumstances (e.g. vegetation, sea surface temperature, ice cover) as well as the physical and dynamical processes of the Earth (including the accuracies of their computations) that drive the atmosphere will be constantly updated to respond to the future climate that during those days is the present climate of the operational models. Robustness of algorithms is commonly validated by sensitivity tests when the algorithms are implemented in NWP models. A more challenging and less analysed robustness challenge for the models, in relative terms, can be extreme events. Extreme events are so rare even among some dozens of ensemble members that ensembles may mismatch the events. Also, at least to some extent, assimilations and dynamical-physical algorithms of the models can be prone to cascading or thresholding effects that may introduce either too weak or too strong intensification of the extreme storms and phenomena associated with them. In the following text paragraphs no detailed general presentation of the uncertainty and robustness of storm hazard predictions is given but we only shortly describe those validation procedures and results that are available of the predictions FMI will provide for the ANYWHERE pilot.

5.9.1 Uncertainties with the predictions in the LUOVA bulletin

The main storm hazard product provided by FMI for all civil protection authorities in Finland is the LUOVA bulletin. It contains information of the predicted wind speeds for the next 1-2(-5) days, their geographical distribution and timing in the form of agreed tables and as free text. Uncertainties can be characterized by words if they are larger than usually but in general the LUOVA forecasts are a synthesis of the deterministic and probabilistic NWP model output (ECMWF IFS, Arome-Harmonie MEPS, HIRLAM) edited by the duty meteorologist. The LUOVA bulletin includes meteorologist's estimates of the hazard levels in the form of the activity levels ("traffic lights") of the civil protection authorities. One of the receiving authorities is the situational awareness centre ISTIKE that is the ANYWHERE Pilot site in Finland. The end users have validated the robustness of LUOVA forecasts in extreme weather events for assessing the performance of the NWP-based predictions (Table 5). The

events are the “Tapani storm (December 2011)” which is a good example of an intense, synoptic-scale low-pressure windstorm that are usually occurring in Finland during the autumn/winter season. In the large-scale weather event (low pressures), the phenomena can be forecasted several days ahead so the timeline of ISTIKE’s preparation is rather long. The shorter time-scale actions are simulated with a historical case “Asta thunderstorm (July 2010)”. Asta thunderstorm was an intense mesoscale convective storm affecting smaller area than Tapani. Because of the typical nature of convective storms, e.g. rapid development, poorly known exact hit location and limited coverage, only very short forecast times can provide reasonable skill of the predicted intense winds. Even then the impacts of the phenomenon are often severe. Also, because the predictability of thunderstorms is so short, the timeline of the actions of ISTIKE is much shorter than in the larger scale events. Both storm cases are very well documented and provide typical weather related risks of the Finnish environment. However, the magnitude of the damage was unusually large in the two cases.

Table 5: End users’ validation scores of the hazard predictions of LUOVA bulletins in extreme weather in Finland.

Indicator	**Description**	**Goal / Typical value**
Predictability of rainfall (h)	The calculation of the uncertainty is based on the forecaster’s edited rainfall forecasts that were compared to the observations in Finland (24h). The validation score used in the evaluation is Stable Equitable Error in Probability Space (SEEPS) introduced by Rodwell et al., (2010).	98h / 96-107 h
Hit rate of warnings of strong winds for 1-2 days	Verification of wind warnings is done with the operational validation system running at FMI. The system compares forecasts and given wind warnings to the observed wind speed.	81 % / 78-85 %
Customer satisfaction of the users of LUOVA-bulletin (scale 1-5)	**Customer satisfaction of the users of LUOVA-bulletin:** • **Overall grade according to the end users: 4.1-4.3** • Quality and accuracy (usually between 4-5) • Necessity (usually between 4-5) • Forecasting time (usually between 4-5) • Clarity of content (usually between 4-5)	3.9 /
Hit rate of the hazards in LUOVA-bulletin	Verification of is done with the operational system running at FMI. It is done based on the observation and model data of FMI and data provided by the rescue service.	73 % / 76-80 %
Predictability of the hazards in LUOVA-bulletins (h)	Verification of is done with the operational system running at FMI. It is done based on the observation and model data of FMI and data provided by the rescue service.	22 h / 26-40 h

Because of the typical nature of convective storms, e.g. rapid development, poorly known exact hit location and limited coverage, only very short forecast times can provide reasonable skill of the predicted intense winds. Even then proper nowcasting of the thunderstorm gusts is difficult, since established analysis tools of Doppler radar –based wind analysis and warning are lacking or are in the phase of R&D (e.g. the SASSE algorithms for the electricity grid in Eastern Finland in WP5). Therefore, the existing nowcasting procedures of the strength of the wind gusts of thunderstorms still utilize the observations once the extreme winds have been measured with ground-based sensors in association with each convective storm. Since the predictability of thunderstorms is so short, the timeline of the actions of ISTIKE is much shorter than in the larger scale events. Both storm cases are documented and provide typical weather related risks of the Finnish environment. Unfortunately, a detailed written meteorological analysis of the predictions (e.g. LUOVA bulletins), the verifying analysis of the real time-space distribution of the hazards and the impacts is not available for this ANYWHERE Deliverable. However, it can be said in general that the magnitude and the impact regions of Tapani were well predicted many days before the storm hit whereas the intensity as well as the exact timing and location of Asta were not well recognized until the storm crossed the Finnish-Russian border and the damage and high wind speed measurements in the Finnish territory became recognized.

Predictions in the LUOVA-bulletin consist of a kind of blending of both hazards and their impacts on the work of civil protection authorities. In the background the hazard predictions contain of course all uncertainties in the existing NWP models and nowcasting tools. In the following we will describe only some of them, not available in the verification routines of the ECMWF IFS process.

5.9.2 Uncertainties of CC-ITN in winter and in the future climate

Convective cells are identified from OPERA- radar mosaics and classified into one of several danger categories using climatology of storm events. For those areas, where OPERA data is not available or because of the quality-issues is not applicable, CC-ITN is utilizing the rainfall intensity estimation based on GLD360 flash density. The lightning data provides important additional information to the radar-based rainfall estimate outside radar coverage and reveals the areas influenced by intense rainfall. Lightning-based estimate can also capture the peak intensities of rainfall, which is an important property from the viewpoint of end users of the ANYWHERE products.

Convective storm identification, tracking and nowcasting (CC-ITN) algorithm is so far developed for convective summer storms, and thus the statistical storm severity classes are determined only for summer storms with rainfall. Algorithm is not applicable for winter storms, because the precipitation rate conversion of the measured reflectivity only to rain is applied for the OPERA data (Saltikoff et al. 2015). For snowfall the conversion is more complex (Rasmussen et al. 2003, von Lerber et al. 2017), and with snowfall both the advection and the vertical profile of reflectivity (VPR) cause much more uncertainty to the precipitation estimate on the ground, especially at ranges beyond 50-100 km, than uncertainties with the proper conversion (Lauri et al. 2012, Koistinen and Pohjola, 2014). The European radar

network consists of partly dual-polarization and partly more traditional single-polarization weather radars, therefore for a uniform rain estimate, the conversion factors are based on single-polarization measurements. The rain estimate converted from reflectivity measurements can overestimate the rain intensity with heavy rainfall as well as with hail. A hail correction utilizing radar-based hail indicator (Waldvogel et al. 1979, Holleman, 2001) is developed to the OPERA data, although not yet implemented.

5.9.3 Robustness of CC-ITN

Severe thunderstorms are much more likely to form in environments with large values of convective available potential energy (CAPE) and deep-tropospheric wind shear (Brooks, 2013) In future climate, the hypothesis indicates that CAPE will increase and the wind shear will decrease. On average CAPE will increase as the surface temperature and boundary layer moisture increases and shear will decrease as the equator-to-pole temperature gradient decreases (Brooks, 2013). Detailed analysis suggests that the CAPE change will lead to more frequent environments favourable for severe thunderstorms, but the strong dependence on shear for tornadoes, particularly the strongest ones, and hail means that the interpretation of how individual hazards will change is open to question. The expectations, how these will change the storm conditions in Europe is not clear. EU-funded 7th program RAIN-project researched the impact of extreme weather events on transport, energy and telecommunication networks and changes in future climate (Groenemeijer P et al. 2016). It was foreseen that the annual frequency of thunderstorms related with severe wind gusts (> 25m/s) increases in central and south-central Europe and the probability of heavy rain increases in many places, except in southwest Europe.

The CC-ITN algorithm is based on measured reflectivity and therefore, there is no restrictions of applying it for future storms (Rossi et al. 2014, Rossi et al. 2015). However, the statistical determination of storm severity is likely to alter and must be determined based on new statistical data. The ability to track the storm cells is preliminary tested and described in deliverable D1.3. In task 2.8. and WP5 the severity estimation compared to severity and real impacts can be evaluated and reported in deliverable D2.5.

6 Summary of main characteristic and uncertainty/robustness facts of selected model/algorithms/tools included on the MH-EWS

Table 8 summarizes the main outputs of the uncertainties and robustness preliminary assessment carried out up to now.

Overall, the large variety of models/algorithms/tools along with a variety of appropriate assessment criteria precluded definition of a common approach for all cases. Instead, we focus here on describing the main facts related to uncertainties and robustness on an individual basis. In general terms, a large effort has been done to quantify different sources of uncertainties. Algorithms/tools have been tested against recent historical events showing an acceptable performance, although it depends on the temporal scale and different geographical contexts. Several models/algorithms/tools have shown high sensitivity to the reliability of the meteorological forecasts. Thus, uncertainties due to this have been reduced with improvement of meteorological models. Consequently, under future climate conditions, the robustness of the algorithm/tools to assess impacts of natural hazards may also depend on the robustness of the climate model scenarios.

Overall, algorithms to forecast the type and intensity of precipitation showed good skills at different temporal scales in forecasting rain, but have difficulties to forecast mixtures of rain and snow. This may have some drawbacks in mountain regions, where a mixture of snow and rain is common, but could be robust in other regions. Flood nowcasting algorithm linked to radar precipitation products were able to reproduce properly flash floods in Catalonia. The robustness of this algorithm depends on the estimation of the return period of the observed and forecasted catchment-aggregated precipitation, which will rely on the performance of the RCM to project precipitation. Thus, region with robust signals in terms of precipitation predictions may benefit most from this tool. The EFAS algorithm reported a good agreement between flood alerts and the number of people affected by floods, especially in central Europe. In fact, the Nash-Sutcliffe efficiency criterion shows that the model has explanatory power for 90% of the analysed catchments (NSE>0). The algorithm seems to perform well for extreme events, regardless of the typology of events and topography, which again suggests robustness. Yet, some inconsistencies remain, especially in southern Europe where impacts and damage of extreme events may increase (Rojas et al., 2013). The storm surge algorithm showed an acceptable result, although it differs between geographic contexts, mostly due to the lack of accuracy in data, but also due to the mesh resolution in areas with complex bathymetry. Besides, storm algorithm reported proper skills for forecasting strong wind warnings within1-2 days ahead. Drought algorithms have been able to forecast current drought conditions in Catalonia, although presumably with some uncertainties for the next months. Under future conditions, the number of days triggering drought conditions may increase in southern Europe. The test is provided against current conditions, which is one of the most extreme events on record; it suggests that the algorithm would be quite robust under CC condition. The UTCI model focuses on weather-induced heatwaves and related health impacts and provided useful

assistance to disaster preparedness plans, since it is able to forecast impacts especially at short-to-medium ranges (up to 4-6 days). The algorithm has also forecast skill at medium range (up to 10 days), especially in coastal areas. Besides, the algorithm seems to be slightly less efficient for heat stress than for cold stress. Also, the model seems to be more robust in the Mediterranean area and northern Europe, where impacts will increase under future CC condition. Finally, wildfires were anticipated with more accuracy at short to medium range and once the ignition as already taken place.

Table 8 Summary of uncertainties and robustness assessment

Models/ algorithms/ tools	Descriptive	Questionnaire	Uncertainties	Robustness	Comments
Extreme precipitation related hazards	The verification is carried out based on 3-hourly observations of present from (SYNOP stations in Europe) at two different lead times 24-48 h and 96-120 h.	Uncertainties of input variables are estimated Model is sensitive to state variables. Uncertainties is not assessed. Model is sensitive to parameter uncertainty. Model has been tested under very extreme conditions The estimation of FAR is on process Uncertainties related to Cascading effect has not been estimated. Uncertainties related to CC not estimated	Rain forecasts are reasonably reliable for both lead times and Rmin settings The Snow forecasts are also reasonably reliable Freezing rain and rain and snow mixed forecasts are not good but show some limited skill, although freezing rain shows better skill in shorter lead times	Robustness assessment based on Symmetric Extremal Dependency Index (SEDI) for different lead times. Freezing rains shows better skill in shorter lead times with a maximum value. Rains and snow show a more stable behave over lead times. Ice pellet forecasts have no skill	Model represent well rains event at different temporal scale, but has difficulties to forecast rains and snow which may affect the suitability to anticipate rain-on-snow events (which are expected to be more probable in mountain regions)
Flash floods	Model based on high-resolution precipitation inputs generated from radar observations and nowcasts	Uncertainties to input variables is estimated (Berenguer et al., 2011; Alfieri et al., 2017). Model is not sensitive to state variables. Model is not sensitive to parameter variability. Not tested to very extreme events. FAR is being assessed.	Model outputs suggest good agreement with real on two flood scenarios in Catalonia.	Robustness require to estimate return period of the observed and forecasted catchment-aggregated precipitation. Consequently, relies on the performance	Appropriate for nowcasting, Potential drawbacks is the static view of channel and not inclusion of the role of human

		Cascading effects are not considered Model has not been tested under CC		of RCM to project precipitation.	
European Flood Awareness System (EFAS)	Continuously monitored and evaluated against its own climatology Quality limited by uncertainties related to hydrological model and meteorological products	Uncertainties of input variables are estimated by perturbing input variables. Model is sensitive to state variables, but uncertainties have not been estimated. Model is sensitive to parameter variability and it is estimated (Zajac et al., 2009; Salamon and Feyen, 2009). Model has been tested under very extreme conditions (Smith et al., 2016). FAR has been estimated (Bartholmes et al., 2009) Uncertainties related to Cascading effect has not been estimated. Uncertainties related to CC has been tested (Based on expected annual damages: Rojas et al., 2013).	Based on Nash-Sutcliffe efficiency criterion, the model shown to have explanatory power for 90% of the catchments (NSE>0). In 32% of the catchments, the model explains over three quarters of the variance of the observed series. Good agreement between the observed and simulated flow statistics in Central Europe. Large discrepancies in the Iberian Peninsula and on the Baltic coasts Correlation between issued floods alarm and affected people reports up to 0.85	Robustness based on ERIC indicator. ERIC performs well identifying a high proportion of the reported flood events, despite the variety of climatology, topography and land use represented by the case study sites. Good match during extreme events (see table 1)	Model seems to perform well under extreme events, regardless s the typology of events and topography. This suggest robustness. Does not include the effect of anthropogenic managing of river corridors and dams. Less accurate in south Europe, where impact and damages of extreme event may increase (see Rojas et al., 2013)
Storm surges	Verification base on water level time series available from the JRC	Uncertainty related to input variables is not estimated Uncertainty related to state variables is estimated by	Based on tidal data, aceptable results were achieved in most of stations.	A assessment carried out in the Baltic sea provide hit rates up to 0.7	Model perform properly over Europe, although there are

	database Performance was evaluated in terms of different efficiency criterion	predefined range Tested against extreme events FAR is assessed Not tested for Cascading effects Not tested under CC conditions			differences for each region. As the algorithm is driven by RCM, the robustness of this model will impact on the robustness of the forecast.
Drought Forecasting	Medium range forecast. Highly dependent of initial condition and seasonal time-scale (related to ocean at atmosphere features)	Uncertainties of input variables are not estimated. Model is sensitive to state variables, but uncertainties have not been estimated. Model has been tested under very extreme conditions (see result on October 2017 over north Iberia Peninsula). FAR has not estimated. Uncertainties related to Cascading effect has not been estimated. Uncertainties related to CC not estimated	Tested on Catalonia Pilot Site (autumn 2017-spring 2018) seasonal droughts: based on forecasting volume of reservoirs. Good signal for October conditions, uncertainty in results over next months. Drought forecasting: good match for October low reservoir levels, uncertainty afterwards. River discharge: not yet evaluated.	Hardly investigated Expected to be robust under CC	The model has been capable to recognize the current drought conditions in a ANYWHERE Pilot Site where the severity of events is projected to increase in the coming decades
Wildfires	Forecasting wildfires is a complex as depend on stochastic processes. Human also play a role. The algorithm is focused on	Algorithm sensitive to input and state variables (Di Giuseppe et al., 2016) Algorithm sensitive to parameter variability Not tested against extreme events FAR is assessed by R	Caliver's methodology returns high level of hits and lower number of misses events at European scale and country level At northern countries	This investigation is still work in progress. Although the algorithm is more robust at short to medium range and once the ignition as	Model perform appropriate at short-medium term.

	determining favourable weather conditions which would allow sustained fire activity.	package (CALIVER) No tested cascading effect neither against CC scenarios	seems to be better to include country-specific danger levels. However, at Mediterranean countries European levels seems more appropriate.	already taken place.	
Weather-induced heatwave and related health impacts	The skill of the UTCI forecasts is based on the quality of the meteorological input, specifically to 10 m wind speed, 2 m relative humidity, 2 m temperature, and solar radiation	Uncertainties of input variables are estimated by perturbing input variables (Pappenberger et al., 2015) Model is sensitive to state variables. Model has been tested under very extreme conditions (Pappenberger et al., 2015). The estimation of FAR is on process Uncertainties related to Cascading effect has not been estimated. Uncertainties related to CC not estimated	Forecasts have skill in the medium range (up to 10 days). Probabilistic UTCI forecasts shows the most skilful, although shows regional variation related to the sensitivity of wind speed, for instance, in coastal areas. Predictability seems to be slightly lower for heat stress than for cold stress	Robustness assessment carried out on Mediterranean area and north Europe suggest that the Anomaly Correlation Coefficient stays above 60% during 4-6 days on average.	The UTCI can be used in daily forecasts and early warnings of extreme weather events to assist disaster preparedness plans
Storms (especially in the Luova bulletins from FMI)	Predictability of extratratropical cyclones is quite good. Predictability of convective storms is so short Luova bulletin report performance,	Model sensitive to input variable Model not sensitive to state variables, but to parameter variability Not tested under very extreme events FAR is not assessed Not tested for Cascading	Given by different indicators of the end users: Predictability of rainfall: Values are within the expected range Accuracy of strong wind warnings 1-2 days: Values are within the		Customer satisfaction based on different indicator provide a value of 3.9/5

	rather than uncertainties as exact metrics.	effects Not tested under CC conditions	expected range Accuracy of LUOVA-bulletin: slightly lower than expected Predictability of LUOVA-bulletins (h): slightly lower than expected

7 References

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

Alfieri, L., Pappenberger, F., Wetterhall, F., Haiden, T., Richardson, D. & Salamon, P. (2014). Evaluation of ensemble streamflow predictions in Europe. J. Hydrol., 517, 913-922, doi:10.1016/j.jhydrol.2014.06.035

Alfieri, L., Thielen, J. (2015). A European precipitation index for extreme rain-storm and flash flood early warning. Meteorological Applications, 22(1), 3-13.

Alfieri, L., M. Berenguer, V. Knechtl, K. Liechti, D. Sempere-Torres, and M. Zappa, (2017). Flash Flood Forecasting Based on Rainfall Thresholds. Handbook of Hydrometeorological Ensemble Forecasting, Q. Duan, F. Pappenberger, J. Thielen, A. Wood, H. L. Cloke, and J. C. Schaake, Eds., Springer Berlin Heidelberg, doi: 10.1007/978-3-642-40457-3_49-1.

Arnal, L., H.L. Cloke, E. Stephens, F. Wetterhall, C. Prudhomme, J. Neumann, B. Krzeminski, F. Pappenberger (2017). Skilful seasonal forecasts of streamflow over Europe? Hydrol. Earth Syst. Sci. Discuss., https://doi.org/10.5194/hess-2017-610.

Barriopedro, D., E. M. Fischer, J. Luterbacher, R. M. Trigo, and R. Garcia-Herrera. (2011). The hot summer of 2010: Redrawing the temperature record map of Europe." Science 332 (6026): 220–224

Bartholmes, J., Thielen, J. & Kalas, M. (2008). Forecasting medium-range flood hazard on European scale. Georisk: Assessment and Management of Risk for Engineered Systems and Geohazards, 2, 181-186, doi:10.1080/17499510802369132.

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

Borga, M., Stoffel, M., Marchi, L., Marra, F., & Jakob, M. (2014). Hydrogeomorphic response to extreme rainfall in headwater systems: flash floods and debris flows. Journal of Hydrology, 518, 194-205.

Bloomfield, J.P., Marchant, B.P. (2013). Analysis of groundwater drought building on the standardized precipitation index approach. Hydrol. Earth Syst. Sci., 17, 4769–4787, doi:10.5194/hess-17-4769-2013

Brooks H.E.,2013, Severe thunderstorms and climate change, In Atmospheric Research, Volume 123 (2013). Pages 129-138, ISSN 0169-8095, doi 10.1016/j.atmosres.2012.04.002.

Buizza, R., and T. N. Palmer, (1998). Impact of ensemble size on ensemble prediction. *Mon. Wea. Rev.*, 126, 2503–2518.

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

Cloke, H. L. & Pappenberger, F. (2008). Evaluating forecasts of extreme events for hydrological applications: an approach for screening unfamiliar performance measures. Met. Apps, 15, 181-197, doi:10.1002/met.58.

Ferro, C., and D. Stephenson, (2011). Extremal dependence indices: Improved Verification measures for deterministic forecasts of rare binary events. *Wea. Forecasting*, 26, 699–713, doi: 10.1175/WAF-D-10-05030.1.

Gascón, E., Hewson, T., Haiden, T., (2017). Improving predictions of precipitation type at the surface: Description and verification of two new products from the ECMWF ensemble. Weather and Forecasting (in press).

Germann, U., Berenguer, M., Sempere-Torres, D., & Zappa, M. (2009). REAL—Ensemble radar precipitation estimation for hydrology in a mountainous region. Quarterly Journal of the Royal Meteorological Society, 135(639), 445-456.

Ghelli, A., A. Garcia-Mendez, F. Prates, and M. Dahoui. (2010). Extreme weather events in summer 2010: how did the ECMWF forecasting systems perform? ECMWF Newsletter (125) 7-11.

Groenemeijer P., Vajda A., Lehtonen I., Kämäräinen M., Venäläinen A., Gregow H., Becker N., Nissen K., Ulbrich U., Morales Nápoles O., Paprotny D., Púčik T. (2016). Present and future probability of meteorological and hydrological hazards in Europe, FP7 RAIN report 608166-D2.5.

Hanley, J. A., and B. J. McNeil, (1982). The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology*, 143, 29–36.

Havenith, G., D. Fiala, K. Blazejczyk, M. Richards, P. Bröde, I. Holmer, H. Rintamaki, Y. Benshabat, and G. Jendritzky, (2012). The UTCI-clothing model.” Int. J. Biometeorol. 56 (3): 461–470.

Heudorfer, B., Stahl, K. (2017). Comparison of different threshold level methods for drought propagation analysis in Germany. Hydrology Research, 48.5, 1311-1326.

Holleman, I., 2001: Hail detection using single-polarization radar. Scientific report 2001/01, Royal Netherlands Meteorological Institute (KNMI)

Ionita M, Lohmann G, Rimbu N. (2008). Prediction of Elbe discharge based on stable teleconnections with winter global temperature and precipitation. Journal of Climate 21: 6215–6226. DOI:10.1175/2008JCLI2248.1.

Ionita M, Dima M, Lohmann G, Scholz P, Rimbu N. (2015). Predicting the June 2013 European flooding based on precipitation, soil moisture and sea level pressure. Journal of Hydrometeorology 16: 598–614. DOI:10.1175/JHM-D-14-0156.1.

Knijff, J. M. V. D., Younis, J. & Roo, A. P. J. D. (2010). LISFLOOD: a GIS-based distributed model for river basin scale water balance and flood simulation. International Journal of Geographical Information Science, 24, 189-212, doi:10.1080/13658810802549154.

Knutti, R., & Sedláček, J. (2013). Robustness and uncertainties in the new CMIP5 climate model projections. Nature Climate Change, 3(4), 369-373.

Koistinen, J. and H. Pohjola, (2014). Estimation of Ground-Level Reflectivity Factor in Operational Weather Radar Networks Using VPR-Based Correction Ensembles, Journal of Applied Meteorology and Climatology 2014 53:10, 2394-2411.

Lauri, T., J. Koistinen, and D. Moisseev, (2012). Advection-Based Adjustment of Radar Measurements. Monthly Weather Review, 140:3, 1014-1022.

Lavaysse, C., Vogt, J., and Pappenberger, F. (2015). Early warning of drought in Europe using the monthly ensemble system from ECMWF, Hydrol. Earth Syst. Sci., 19, 3273-3286, doi:10.5194/hess-19-3273-2015.

von Lerber, A., D. Moisseev, L.F. Bliven, W. Petersen, A. Harri, and V. Chandrasekar. (2017). Microphysical Properties of Snow and Their Link to Ze–S Relations during BAECC 2014. Journal of Applied Meteorology and Climatology, 56:6, 1561-1582.

Magnusson, L., T. Haiden, and D. Richardson, (2014). Verification of extreme weather events: Discrete predictands. ECMWF Tech. Memo., 29 pp.

Mason, I., (1982). A model for assessment of weather forecasts. Aust. Meteor. Mag, 30, 291–303.

Mason, S. J., and N. E. Graham, (1999). Conditional probabilities, relative operating characteristics, and relative operating levels. Wea. Forecasting, 14, 713–725.

McKee, T.B., N.J. Doesken, J. Kleist, (1993). The relationship of drought frequency and duration to timescales, paper presented at 8th Conference on Applied Climatology, Am. Meteorol. Soc., Anaheim, Calif.

Meißner D, Klein B, Ionita M. (2015). Towards a seasonal forecasting service for the German waterways—requirements, approaches, potential products. Presentation at the HEPEX workshop on seasonal hydrological forecasting, 21-23rd September 2015, Norrköping, Sweden (http://hepex.irstea.fr/wp-content/uploads/2015/08/06_Dennis_Meissner_oral_HEPEX_BfG.pdf, accessed 31 October 2017).

Murphy, A. H., and R. L. Winkler, (1977). Can weather forecasters formulate reliable probability forecasts of precipitation and temperature. *Natl. Weather Dig.*, 2, 2–9.

Pegram, G. G. S., and A. N. Clothier, (2001). High resolution space-time modelling of rainfall: the "String of Beads" model. Journal of Hydrology, 241, 26-41.

Pappenberger, F., Bartholmes, J., Thielen, J., Cloke, H. L., Buizza, R. & de Roo, A. (2008). New dimensions in early flood warning across the globe using grand-ensemble weather predictions. *Geophys. Res. Lett.*, 35, doi:10.1029/2008gl033837

Pappenberger, F., Cloke, H. L., Parker, D. J., Wetterhall, F., Richardson, D. S. & Thielen, J. (2016). The monetary benefit of early flood warnings in Europe. Environmental Science & Policy, 51, 278-291, doi:10.1016/j.envsci.2015.04.016.

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

Prudhomme C. (2015). Operational seasonal hydrological forecasting in the UK. Presentation at the HEPEX workshop on seasonal hydrological forecasting, 21–23rd September 2015, Norrköping, Sweden (http://hepex.irstea.fr/wp-content/uploads/2015/08/ 07_HydrologicalOutlookUK_ 21Sep 2015.pdf, accessed 31 October 2017).

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

Quintero, F., D. Sempere-Torres, M. Berenguer, E. Baltas, (2012). Analysis of the uncertainty associated with rainfall and model parameter estimation to flow simulations, Journal of Hydrology, 460-461, 90-102.

Rasmussen, R., M. Dixon, S. Vasiloff, F. Hage, S. Knight, J. Vivekanandan, and M. Xu, (2003). Snow Nowcasting Using a Real-Time Correlation of Radar Reflectivity with Snow Gauge Accumulation, Journal of Applied Meteorology, 42:1, 20-36.

Rodwell M. J. et al., (2010). A new equitable score suitable for verifying precipitation in NWP. Quart. J. Roy. Met. Soc., 136, 1344-1363.

Roebber, P., (2009). Visualizing multiple measures of forecast quality. *Wea. Forecasting*, 24, 601– 608, doi:10.1175/2008WAF2222159.1.

Rossi, P. J., V. Hasu, J. Koistinen, D. Moisseev, A. Mäkelä, E. Saltikoff, (2014). Analysis of a statistically initialized fuzzy logic scheme for classifying the severity of convective storms in Finland. Met. Apps., 21, 656–674.

Rossi, P.J., V. Chandrasekar, V. Hasu, D. Moisseev, (2015). Kalman filtering-based probabilistic nowcasting of object-oriented tracked convective storms, J. Atmos. Oceanic Tech., 32, 461–477.

Saltikoff E., Lopez P., Taskinen A. and Pulkkinen S., (2015). Comparison of quantitative snowfall estimates from weather radar, rain gauges and a numerical weather prediction model. Boreal Env. Res. 20: 667–678.

Shukla, S., Wood, A.W., (2008). Use of a standardized runoff index for characterizing hydrologic drought. Geophys. Res. Lett, 35, L02405, doi:10.1029/2007GL032487.

Smith, P., Pappenberger, F., Wetterhall, F., Thielen, J., Krzeminski, B., Salamon, P., Muraro, D., Kalas, M. and Baugh, C. (2016). On the operational implementation of the European Flood Awareness System (EFAS), Report 778, European Centre for Medium-Range Weather Forecasting, http://www.ecmwf.int/en/elibrary/355 16337-operational-implementation-european-flood-awareness-system-efas.

Stanski, H. R., L. J. Wilson, and W. R. Burrows, (1989). Survey of common verification methods in meteorology. WMO World Weather Watch Tech. Rep. WMO TD 358.

Tallaksen, L.M,. Van Lanen, H.A.J., (Eds.) (2004). Hydrological Drought. Processes and Estimation Methods for Streamflow and Groundwater. Developments in Water Science, 48, Elsevier Science B.V., 579 pg.T

Taylor, K.E. (2001). Summarizing multiple aspects of model performance in a single diagram. J. Geophys. Res., 106, 7183-7192, 2001

Thielen, J., Bartholmes, J., Ramos, M.-H. & de Roo, A. (2009). The European Flood Alert System - Part 1: Concept and development. Hydrology and Earth System Sciences, 13, 125-140, doi:10.5194/hess-13-125-2009.

Van Wagner, C., and Coauthors, (1987). Development and structure of the Canadian forest fire weather index system, Vol. 35. Canadian Forestry Service, Headquarters, Ottawa.

Van Wagner, C. E., and Coauthors, (1974). Structure of the Canadian forest fire weather index. Environment Canada, Forestry Service.

Vitart, F. (2014). Evolution of ECMWF sub-seasonal forecast skill scores, Q. J. Roy. Meteor. Soc., Part B, 114, 1889–1899.

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

Vitolo, C., F. D. Giuseppe, and M. D'Andrea, (2017b). caliver: an r package for calibration and verification of forest fire gridded model outputs, PlosOne (accepted for publication).

Vousdoukas, M. I., Voukouvalas, E., Annunziato, A., Giardino, A., & Feyen, L. (2016). Projections of extreme storm surge levels along Europe. Climate Dynamics, 47(9-10), 3171-3190.

Waldvogel, A., B. Federer, and P. Grimm, (1979). Criteria for the detection of hail cells. Journal of Applied Meteorology, 18, 1521–1525.

Wanders, N., Wada, Y., Van Lanen, H.A.J. (2015). Global hydrological droughts in the 21st century under a changing hydrological regime. Earth Syst. Dynam. 6: 1–15, doi:10.5194/esd-6-1-2015.Wilks, D. S., 1995: *Statistical Methods in the Atmospheric Sciences*, International Geophysics Series, Vol. 59. Academic Press, 407 pp.

Zajac, Z., Zambrano-Bigiarini, M., Salamon, P., Burek, P., Gentile, A., and Bianchi, A. (2013). Calibration of the lisflood hydrological model for Europe - calibration round 2013. Technical report, Joint Research Centre, European Commission, 2013.