



Project number: 101101229

Call: UCPM-2022-PP

Early warning Demonstration of pan -European RAinfall-induced impacts

Historical Evaluation of EDERA River Flash Flood Forecasts in the Pilot Sites

Deliverable 4.2

10th January 2025

The content of this report represents the views of the author only and is his/her sole responsibility. The European Commission does not accept any responsibility for use that may be made of the information it contains.



Funded by
European Union
Civil Protection
and Humanitarian Aid



DOCUMENT INFORMATION

Title	Historical evaluation of the EDERA River Flash Flood Forecasts in the Pilot Sites
Lead Author	Calum Baugh
Contributors	Juan Pereira Colonese, Christel Prudhomme, Dimitar Tasev
Type	[R — Document, report]
Dissemination Level	[PU — Public]
Document Reference	Baugh, C., J. Pereira Colonese, C. Prudhomme, D. Tasev, 2025: Historical evaluation of the EDERA River Flash Flood Forecasts in the Pilot Sites. EDERA Deliverable D4.2, 32 pp.

VERSION HISTORY

Version	Author	Organisation	Date	Comments
1.0	C. Baugh	ECMWF	30/04/2024	Evaluation of period 2021-2023, a revised document covering 2019-2023 will be produced when forecast data for 2019-2020 become available
2.0	C. Baugh	ECMWF	10/01/2025	The evaluation was extended to cover the five year period from 2019-2023

Table of contents

Executive summary	1
1 Introduction and scope	2
2 Datasets used for Evaluation.....	2
2.1 Observations	3
2.1.1 Media reports of Riverine Flash Flood Impact	3
2.2 River Flash Flood Forecasts	6
3 Evaluation Methodology	7
3.1 Diagnostics.....	7
3.2 Analysis of Notifications	7
3.3 Evaluation of Forecast Performance against Media Reports of Flash Flood Impacts....	7
4 Results.....	8
4.1 Diagnostics.....	8
4.1.1 Pan-European Results	8
4.1.2 Historical Analysis of Flash Flood Notifications in Finland	13
4.1.3 Historical Analysis of Flash Flood Notifications in Andalusia	16
4.2 Evaluation against Media Reports of Flash Flood Impacts	19
4.2.1 Finland	19
4.2.2 Spain and Portugal.....	20
5 Discussion	25
6 Conclusions	26
References	27

List of tables

No table of figures entries found.

List of figures

Figure 1 Monthly distribution of unique meteorological events associated with river flash flooding in Spain and Portugal between 2019-2023.....	4
Figure 2 Total number of unique meteorological events associated with river flash flooding in each NUTS level 3 administration region in Spain and Portugal between 2019-2023.	5
Figure 3 Monthly distribution of unique meteorological events associated with river flash flooding in Finland between 2019-2023 and the total number of events in each river catchment.....	6
Figure 4 Number of forecast production days when each subcatchment had a forecasted impact level ≥ 1 for each of the four summary lead time periods between 2019-2023.	9
Figure 5 Number of forecast production days when each subcatchment had a forecasted impact level ≥ 1 in each season, for each of the four summary lead time periods.....	11
Figure 6 Number of forecast production days when each subcatchment had a forecasted impact level in each of the four impact categories.	12
Figure 7 Number of notifications per week which would have been disseminated to SYKE according to their notification criteria.	14
Figure 8 Total number of forecast production days when conditions in each basin met the criteria to generate a notification for SYKE.....	15
Figure 9 Seasonal distribution of the number of forecast production days between 2019-2023 when the forecast in each basin satisfied the notification criteria provided by SYKE.	16
Figure 10 Number of notifications per week which would have been disseminated to AMAYA according to their notification criteria for 0-6 hours lead time (left) and 7-24 hours lead time (right).	17
Figure 11 Total number of forecast production days when conditions in each basin met the criteria to generate a notification for AMAYA for 0-6 hours lead time (left) and 7-24 hours lead time (right).	18
Figure 12 Seasonal distribution of the number of forecast production days between 2019-2023 when the forecast in each basin satisfied the notification criteria provided by AMAYA for 0-6 hours lead time (top) and 7-48 hours lead time (bottom).....	18
Figure 13 Hits, misses and false alarms from the evaluation of the EDERA forecasts applying different impact level thresholds against flash flood observations in Finland between 2019-2023 at different lead time ranges.	19
Figure 14 Catchment distribution of hits, misses and false alarms between 2019-2023 in Finland for 7-24 hours lead time applying a minimum impact level threshold of 3.	20
Figure 15 Recall score (hits / (hits + misses)) across Spain and Portugal between 2019-2023 for different lead time ranges when applying different minimum impact level thresholds.	21
Figure 16 Precision score (hits / (hits + false alarms)) across Spain and Portugal between 2019-2023 for different lead time ranges when applying different minimum impact level thresholds....	22

Figure 17 F(beta) score across Spain and Portugal between 2019-2023 for different lead time ranges when applying different minimum impact level thresholds.....	23
Figure 18 Hits, misses and false alarms over Spain and Portugal between 2019-2023 at 0-6 hours (top left), 7-24 hours (top right), 25-48 hours (bottom left) and 49-120 hours (bottom right) lead time ranges when applying different impact level thresholds.	24
Figure 19 NUTS 3 regional distribution of hits, misses and false alarms between 2019-2023 in Spain and Portugal for 0-6 hours lead time applying a minimum impact level threshold of 3 (top) and 7-24 hours and 25-48 hours lead time applying a minimum impact level threshold of 2 (middle and bottom respectively).....	25

Executive summary

In this report the EDERA river flash flood forecasts were generated during the period 2019-2023 to analyse and evaluate their performance at the pan-European scale and in both EDERA study sites (Finland and Spain/Portugal). Across Europe, the forecasts were analysed to show how often flash flood impacts were forecast in each subcatchment and the distribution according to the season and forecast impact severity level. In each study site, the forecasts were analysed according to the notification criteria provided by the SYKE and AMAYA stakeholders and were also evaluated against media observations of flash flood impacts.

At the pan-European scale, the forecasts for the most severe flash flood impacts were confined to lead times less than 24 hours. At longer lead times the increased forecast uncertainty resulted in lower forecasted impact severities. Results showed a possible issue related to the masking of snowfall at longer lead times which resulted in over-warning of flash flood impacts, this merits further investigation in future work.

Within the Finland study site, it was only possible to evaluate the forecasts against two observed events which showed a limited performance of the forecasts beyond 24 hours lead time. These findings would suggest that the criteria specified by SYKE to receive notifications should be revised to only issue notifications where the impact level exceeds a value of three up to a maximum lead time of 24 hours. In Spain and Portugal, good forecast performance was achieved up to 48 hours lead time, the original notification criteria specified by AMAYA were found to give the best results. Differences in the number of notifications generated between 0-6 and 7-48 hours lead time could suggest differences in the forecast rainfall patterns between the nowcast and NWP lead time ranges and merits investigation in future work.

Future work should investigate improving the performance of the forecasts at lead times beyond 48 hours, potentially by different methods to calculate the flash flood probability which consider increased spatial uncertainty in the forecast. Offering additional criteria, such as forecast probability and affected population, for stakeholders to configure when setting up notifications could improve this service and improve the integration of the notifications into stakeholders' decision making workflows when responding to natural hazards.

1 Introduction and scope

Encouraging project stakeholders to incorporate new forecast products, such as those from EDERA, into their workflows can be aided by demonstrating the performance of the products over a longer time period. Previous forecasts during this period can be evaluated against observations of past natural hazard impacts to compute the skill, the results can be communicated to stakeholders to demonstrate the confidence they can place in the forecasts. Results from a skill assessment will also aid the establishment of the notification service offered by EDERA which allows stakeholders to specify the criteria for receiving notifications. Evaluating the performance of the forecast products over a longer time period will give more information to each stakeholder when deciding upon their notification criteria.

In this report, the EDERA river flash flood products, which forecast the risk of impacts from flash flooding in smaller rivers summarised over subcatchments, were generated each hour during the period 2019-2023. Using the results, firstly a pan-European diagnostic evaluation was performed to analyse general characteristics of the forecasts over the entire continent. The total number of days each subcatchment was forecasted to have flash flood impacts was calculated. This was repeated to break the count down seasonally and by each of the four impact severity levels. The results would give an overview of spatial and temporal trends in the forecast products, for example if some regions were highlighted during particular times of the year or at certain impact levels.

Next, the criteria for sending notifications to the AMAYA and SYKE partners in Andalusia and Finland (respectively) were analysed by counting the total number of notifications which would have been sent to these partners between 2019-2023. The temporal and spatial distribution of these notifications were analysed to determine a reasonable number of notifications were generated in the correct areas at the correct times of year.

Finally, an evaluation of the forecasts against reports of impacts associated with riverine flash flooding, obtained from the media and EDERA project stakeholders, was performed separately in the Spain/Portugal and Finland study sites. The evaluation would analyse the ability of the forecasts to balance maximising the number of successfully forecasted events (hits) whilst minimising the number false positive forecasts (false alarms). This was done by applying different threshold values of the forecasted impact severity level, which ranged from 1 (lowest impacts) to 4 (most severe impacts) and analysing different forecast lead times. Impact severity level and forecast lead time are options which stakeholders can define for receiving notifications, the results from this evaluation can aid stakeholders to optimise their notification criteria.

The results from each of the above analyses would demonstrate the performance of the EDERA forecast products during previous flash flood events. This will enable stakeholders to make a more informed decision about the ability of the forecasts to predict events within their area of interest and how to integrate them into their workflows.

2 Datasets used for Evaluation

In each of the study sites, observations of riverine flash floods during the evaluation period were obtained from media reports of impacts resulting from riverine flash floods. The EDERA flash flood impact forecasts were generated over the 2019-2023 period.

2.1 Observations

2.1.1 Media reports of Riverine Flash Flood Impact

Media reports can provide a good source of observations regarding riverine flash flood impacts, each report can provide the locations affected by flooding and the day on which the flooding occurred. These reports can also provide observations in areas which may not be covered by in-situ monitoring of river discharge or water level. However, media reports are often only produced when impacts are reported in urban areas, meaning that impacts in rural areas can sometimes be under-represented.

In this study, media reports of flash flood impacts were obtained from the website FloodList.com, which collates media reports of flood impacts globally. For each reported flood event, FloodList.com provides information, via an API, regarding the locations where flooding was reported, the date on which the flooding occurred, what impacts were reported, such as the number of people affected or buildings damaged and what type of flooding occurred (e.g. flash flooding, dam breach, large scale riverine flooding). Therefore, reported flood events in Spain, Portugal and Finland between 2019-2023 were downloaded from FloodList.com using its API. The downloaded data were filtered to retain only events where flash flooding was reported as being the main flood type. However, flash flooding is defined by FloodList.com as being caused by either the rapid rise of water within small rivers or the rapid ponding of water in urban areas. This study is focused upon evaluating the forecast products associated only with flash flooding in small rivers. Therefore, each reported flash flood event obtained from FloodList.com was further investigated by analysing the original media reports to see if the flash flooding was driven by either riverine or urban processes. Only the reports where it was confirmed that the flash flooding was driven by riverine processes were retained.

In Finland, there were no reports of riverine flash flooding from FloodList.com. Therefore, project partners at FMI and SYKE were consulted to obtain any reports they may have of riverine flash flood events.

Overall, across the entire pilot site there were 219 reported locations affected by riverine flash floods between 2019-2023. Many of the reported locations are often caused by the same meteorological storm system, taking this into consideration there were 41 unique meteorological events which caused riverine flash flood impacts. Most of the reports were located within Spain and Portugal, only seven reports were available within Finland. The lack of reports in Finland is potentially due to its hydro-morphological context, it is largely flat, its rivers pass through many small lakes which attenuate the downstream propagation of a floodwave, and during winter much of the water is stored within the snowpack which is gradually released through melting in the spring. These conditions would not be conducive to generating flash floods in smaller rivers.

In Spain and Portugal most river flash flood events occurred the autumn and winter seasons (Figure 1), these could be associated with a cluster of events in the north west of the study site (Figure 2) which could be caused by extreme rainfall associated with synoptic frontal systems from the Atlantic ocean. Events were reported in many regions of the study site, many events occurred around Madrid, potentially due to the greater population exposure to flash flooding. NUTS 3 (Nomenclature of territorial Units for Statistics) regions were used to summarise the total number

of events because these are the spatial units for which EDERA notifications are issued to the EDERA project partner AMAYA.

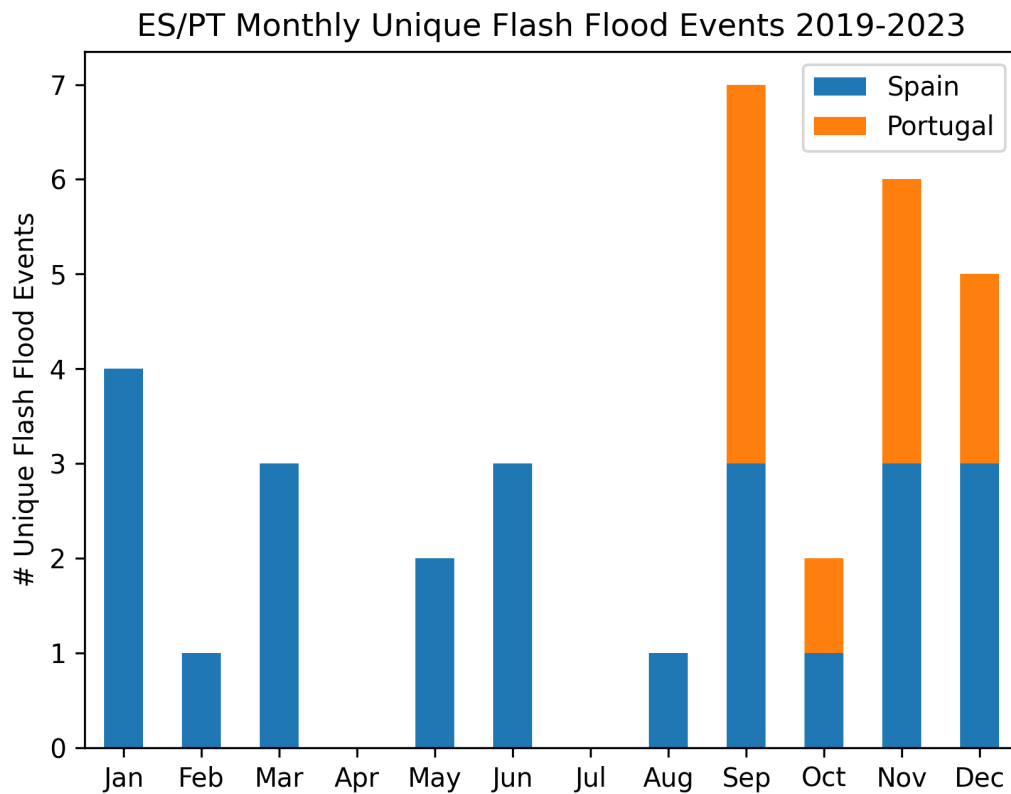


Figure 1 Monthly distribution of unique meteorological events associated with river flash flooding in Spain and Portugal between 2019-2023.

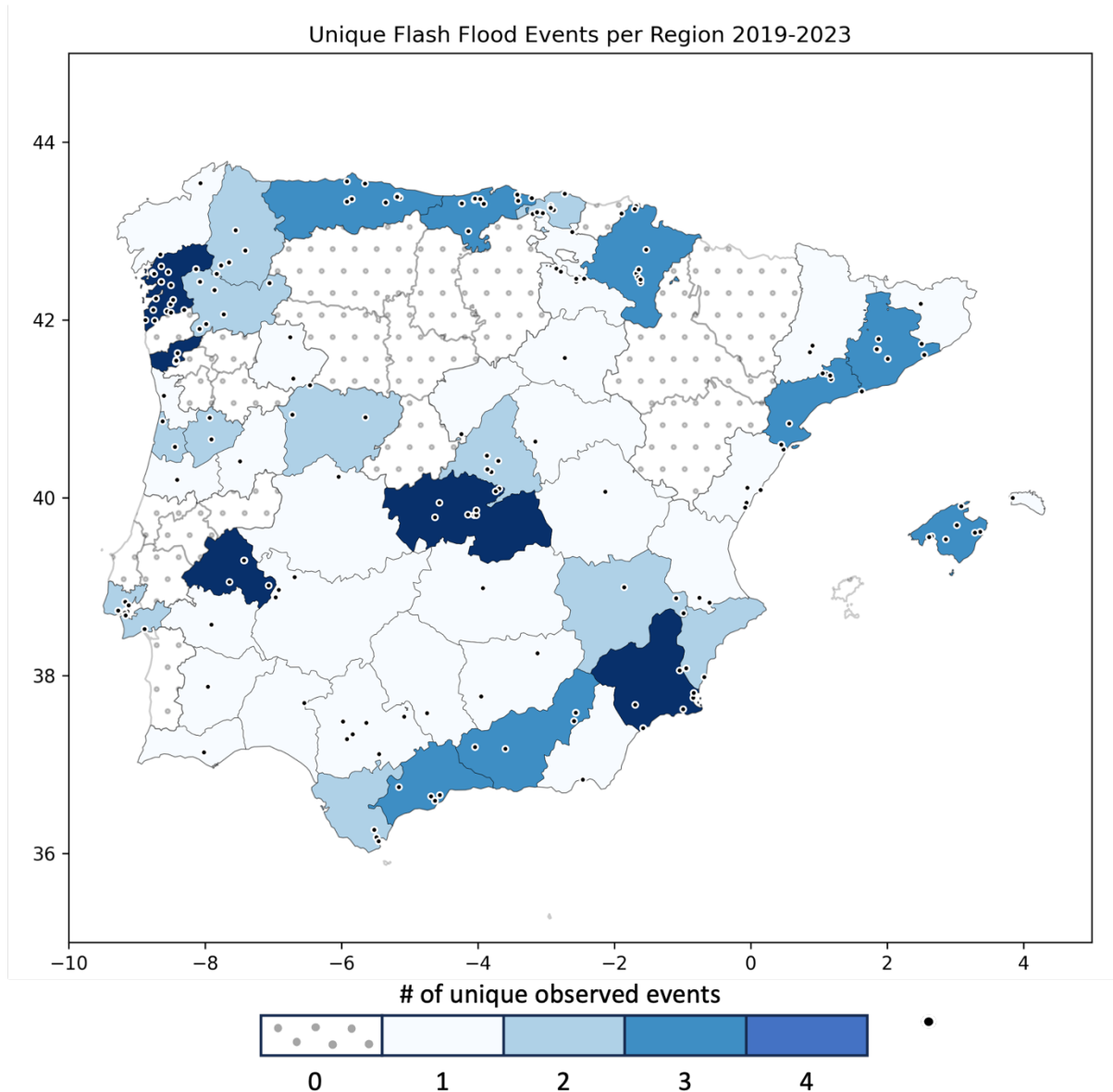


Figure 2 Total number of unique meteorological events associated with river flash flooding in each NUTS level 3 administration region in Spain and Portugal between 2019-2023.

In Finland, the two unique meteorological flash flood events both occurred in 2023, one in September, the other in October (Figure 3). Both events occurred in the north west of the country in the coastal catchments of the Gulf of Bothnia and in the Kemijoki catchment. Catchments are used to summarise the total number of events in Finland as these are the spatial units for which EDERA notifications will be issued to the Finnish partner SYKE.

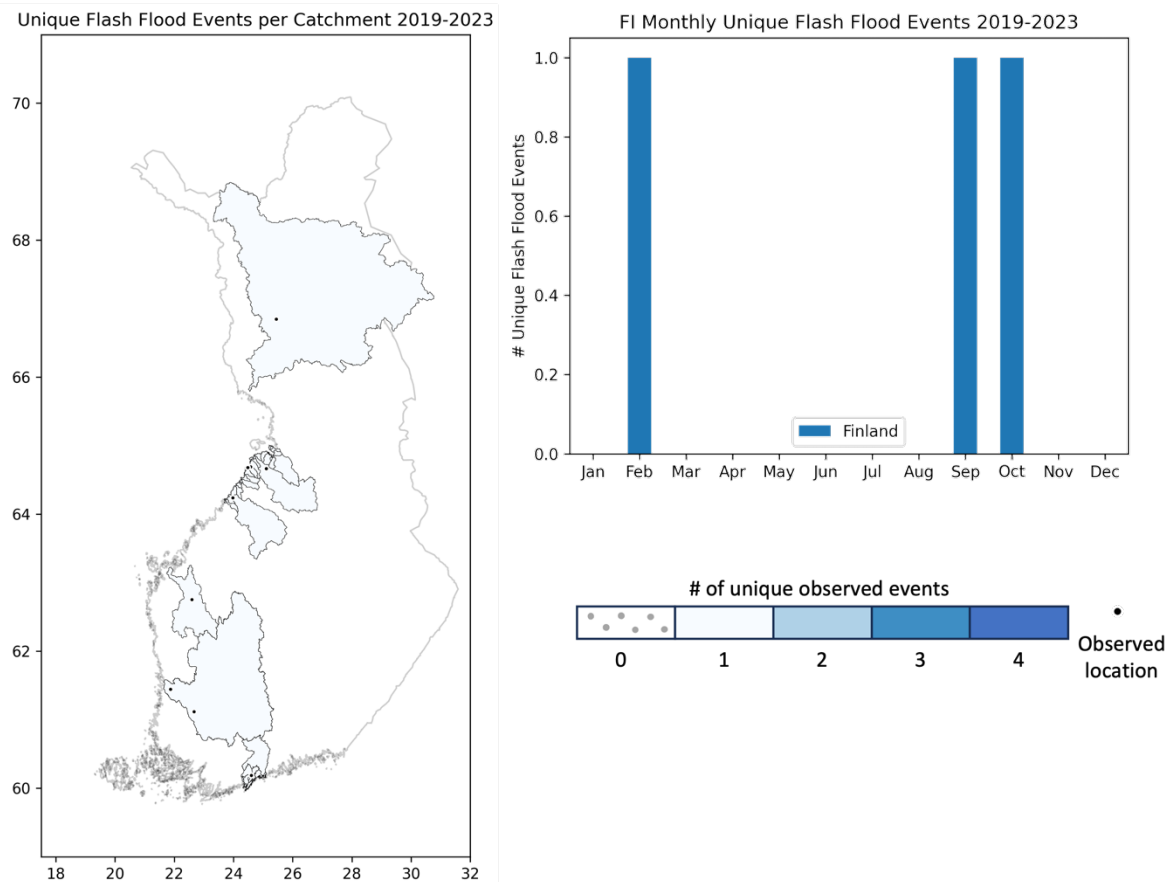


Figure 3 Monthly distribution of unique meteorological events associated with river flash flooding in Finland between 2019-2023 and the total number of events in each river catchment.

2.2 River Flash Flood Forecasts

The EDERA river flash flood forecast products were generated for the 2019-2023 evaluation period using the same methodology to generate the real-time forecasts. NWP (Numerical Weather Prediction) precipitation forecasts were obtained from historical ECMWF ensemble IFS (Integrated Forecast System) rather than reforecasts as the latter has fewer ensemble members. OPERA radar composites [1] during this period were extracted from the data archived at ECMWF, the same rain gauge derived bias correction factor which is applied to the operational EDERA forecasts [2] was applied to each forecast during the evaluation period. The forecasts were generated each hour during each day of the evaluation period, resulting in a total of 8,760 forecasts. For each forecast, the catchment summary product was generated for four lead time ranges (0-6, 7-24, 25-48 and 49-120 hours), which is consistent with the real-time products generated during the EDERA demonstration. This product shows the forecasted flash flood impact level in each subcatchment, which were defined from the 1 arcminute digital elevation model where the time of concentration was less than 24 hours. In each subcatchment, the severity of flash flood impacts over each lead time range is forecasted as one of four levels, a value of 1 being minor impacts and 4 being severe impacts. The data in these catchment summary products were used in the evaluation against observations.

3 Evaluation Methodology

3.1 Diagnostics

A first analysis was performed at the pan-European scale by computing diagnostics including the total number of forecast production days between 2019-2023 when each subcatchment was forecasted to have an impact level of one or greater for each of the four separate lead time ranges for which the catchment summary layers are produced (0-6, 7-24, 25-48 and 49-120 hours). For the final lead time range (49-120 hours), which covers the final three-day period of the total 5 day forecast lead time, if a subcatchment is highlighted during this period it is only counted in the results below as being highlighted for one day, as the analysis counts the number of forecast production days when this subcatchment was highlighted during each lead time range. During each day, 24 separate forecasts are produced, these were aggregated to one daily forecast by taking the maximum forecasted impact level in each subcatchment. This analysis would show if there were any spatial patterns regarding the subcatchments where impacts were forecasted and how this changed with forecast lead time.

The analysis was repeated but with separate counts for each season to identify seasonal patterns in the forecasting of flash flood impacts. Finally, the number of forecast days was counted when the forecasted impact level in each subcatchment was equal to one of the four impact categories.

3.2 Analysis of Notifications

During the demonstration phase of EDERA, the project partner AMAYA and the project stakeholder SYKE have requested to receive real time notifications of forecasted flash flood events in their areas of interest, Andalusia in southern Spain (AMAYA) and Finland (SYKE). Both institutions could specify the criteria to receive notifications, these included the area of interest, lead time range and minimum forecast impact severity level. The criteria specified by both institutions were applied to the EDERA forecasts generated between 2019-2023 to analyse how many notifications would have been issued to each institution over this period.

3.3 Evaluation of Forecast Performance against Media Reports of Flash Flood Impacts

Media reports of flash flood impacts, described above in section 2.1.1, were evaluated separately in the Spain/Portugal and Finland pilot sites. Firstly, the media reports were aggregated for each unique meteorological event onto NUTS 3 regions in Spain and Portugal and river catchments in Finland. These spatial units were selected because they were requested by the AMAYA and SYKE institutions respectively as the spatial unit for which they would receive notifications. Each unique meteorological event often covered multiple days, however the specific day on which each impact observation occurred within the period of the meteorological event was not recorded. Therefore, all forecasts over the same time period as the observation were aggregated by recording the maximum impact level predicted in each subcatchment. Then, for each NUTS 3 (Spain/Portugal) or catchment (Finland) region within the study site, the aggregated forecasts were analysed to see if one or more subcatchments within the region exceeded a specified impact level threshold. If one or more subcatchments satisfied this criterion and impact observations were also reported in the same region, this region was recorded as having a successful forecast i.e. a 'hit'. If the forecast satisfied the criterion but no impact observations were recorded, the region was recorded as a

false alarm; and if impacts were observed in a region but no forecasted subcatchments satisfied the impact level threshold criterion the region was recorded as a miss. This process was repeated for four impact level thresholds (1 to 4, i.e. yellow to purple as displayed on the web viewer, where each value represents low, medium, high and severe risk of flash flood impacts respectively) for each of the four lead time summary ranges used in the catchment summary products displayed on the EDERA web viewer (0-6, 7-24, 25-48 and 49-120 hours). The evaluation was only performed during dates when flash flood impacts were reported, outside of these dates it was not considered possible to evaluate the forecasts as flood impacts may have occurred but not have been reported in the media, this would result in unfairly penalising the forecasts for producing false alarms.

The resulting contingency table of hits, misses and false alarms associated with each lead time range and each impact level threshold were used to compute the following evaluation scores:

$$\text{Recall} = \frac{\text{Hits}}{\text{Hits} + \text{Misses}}$$
$$\text{Precision} = \frac{\text{Hits}}{\text{Hits} + \text{False Alarms}}$$
$$f(\beta) = \frac{(1 + \beta^2) * \text{Hits}}{(1 + \beta^2) * \text{Hits} + \beta^2 * \text{Misses} + \text{False Alarms}}$$

Where β is a parameter governing the weighting towards false alarms, by default its value is set to 1.0

The recall score (also known as the hit rate or probability of detection) shows what proportion of the observed events were successfully forecasted, however this does not consider the role of false alarms. The precision score (also known as the success ratio) considers false alarms and shows what proportion of the forecasted events were observed, it penalises an imprecise forecast which over-warns. The $f(\beta)$ score combines both aspects of recall and precision to give the trade-off between the two, this is important in flash flood forecasting where reducing often large numbers of false alarms must be traded-off against reducing the number of hits.

4 Results

4.1 Diagnostics

4.1.1 Pan-European Results

Using the historical forecast data between 2019-2023, a first analysis counted the number of days during this period that each subcatchment had a forecasted impact level of 1 (yellow i.e. low risk of impacts) or above. The analysis was conducted separately for each forecast lead time range for which the catchment summary products are produced (0-6h, 7-24h, 25-48h and 49-120h). Across all forecast lead times, subcatchments were highlighted most frequently in areas such as the Alps, western Scandinavia, northern UK, Iceland and the mountainous areas of the Balkans (Figure 4). At longer lead times, the frequency of subcatchments highlighted in the Alps, Iceland and western Scandinavia is higher than in other areas. These are areas of higher orography which can enhance rainfall within the NWP forecast, which is the only source of meteorological forcings at these lead times. The orographic enhancement may trigger more events in such areas relative to others where the NWP may trigger extreme rainfall less frequently. At the same longer lead times, the frequency of subcatchments being highlighted is higher across the whole domain. This is likely

because of the longer periods covered by the longer lead time ranges, the final period covers three days compared to six hours for the first range.

Number of Days Subcatchment Impact Level ≥ 1 : 2019 - 2023

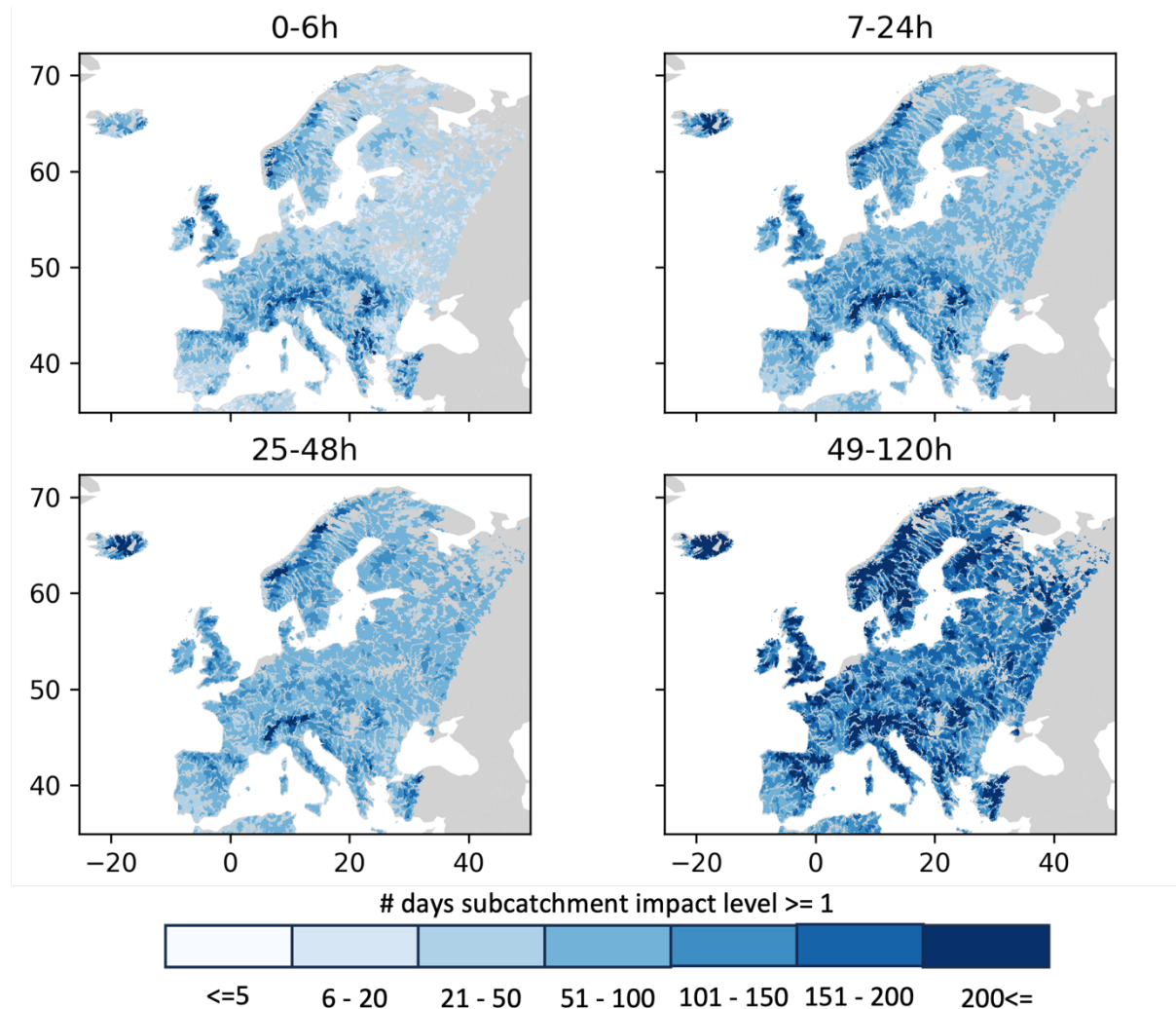


Figure 4 Number of forecast production days when each subcatchment had a forecasted impact level ≥ 1 for each of the four summary lead time periods between 2019-2023.

The total number of days during each season that each subcatchment had a forecasted impact level of 1 (yellow i.e. low risk of impacts) or above was analysed. Results show that the summer season had the highest number of days when subcatchments were highlighted as being at risk of flash flood impacts, autumn having the next highest frequency (Figure 5). During the winter, the number of days when subcatchments were highlighted is lower than other seasons, which is expected given that snowfall will dominate in eastern and northern parts of the domain, whilst in western Europe rainfall is more typically associated with large scale frontal systems which are not always associated with flash flooding. However, during the winter at longer lead times (>24 hours) the results show higher frequencies in western Scandinavia, Iceland, the Alps. The opposite would be expected because most of the precipitation in these areas during winter would fall as snow

hence the frequencies should be lower. This could be related to the process to mask snow areas within the forecast, at shorter lead times snow areas are masked using an NWP forecast probability of snow, however this is not used at longer lead times. These results suggest that an improved masking of snow at the medium range is required.

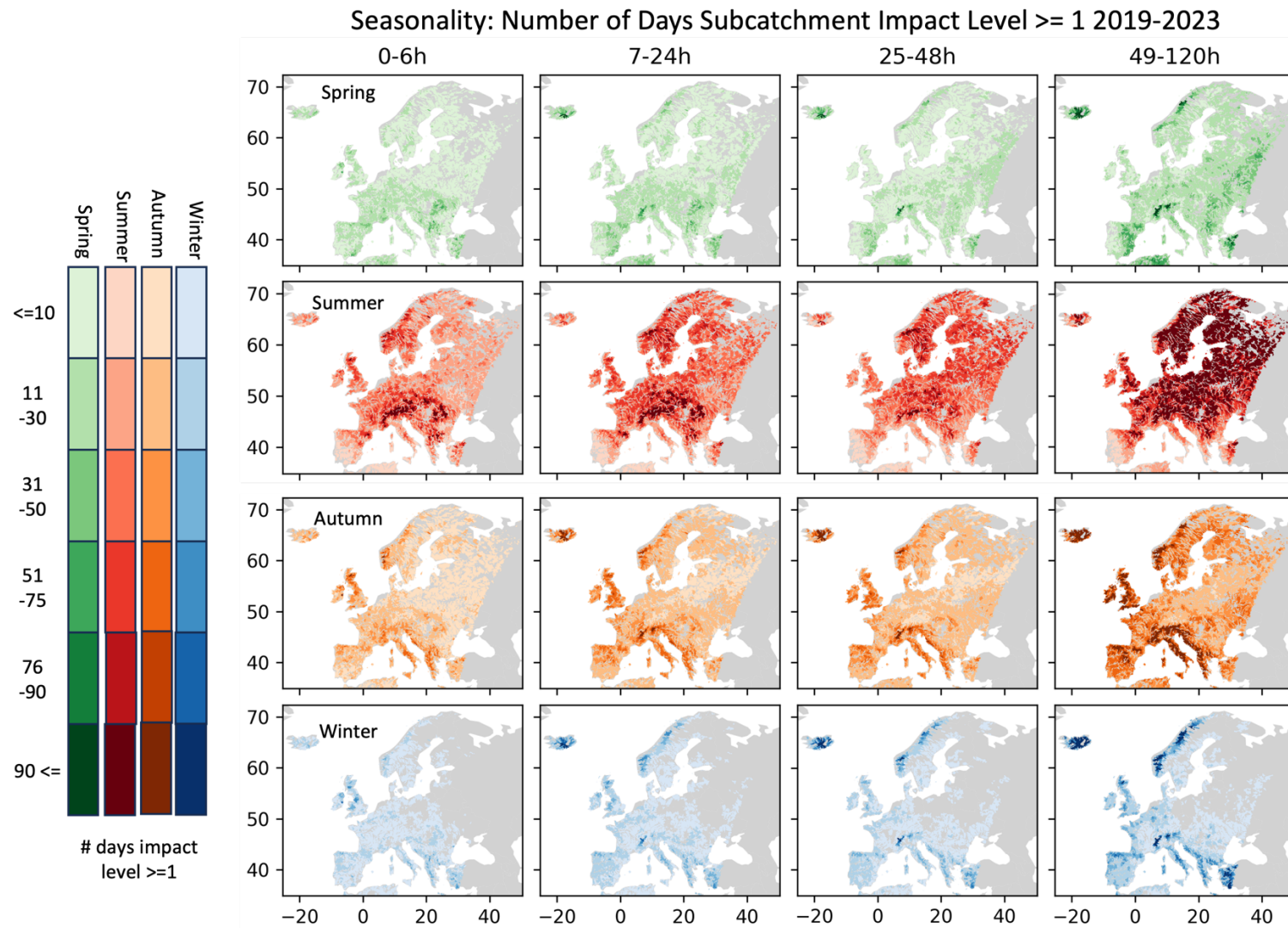


Figure 5 Number of forecast production days when each subcatchment had a forecasted impact level ≥ 1 in each season, for each of the four summary lead time periods

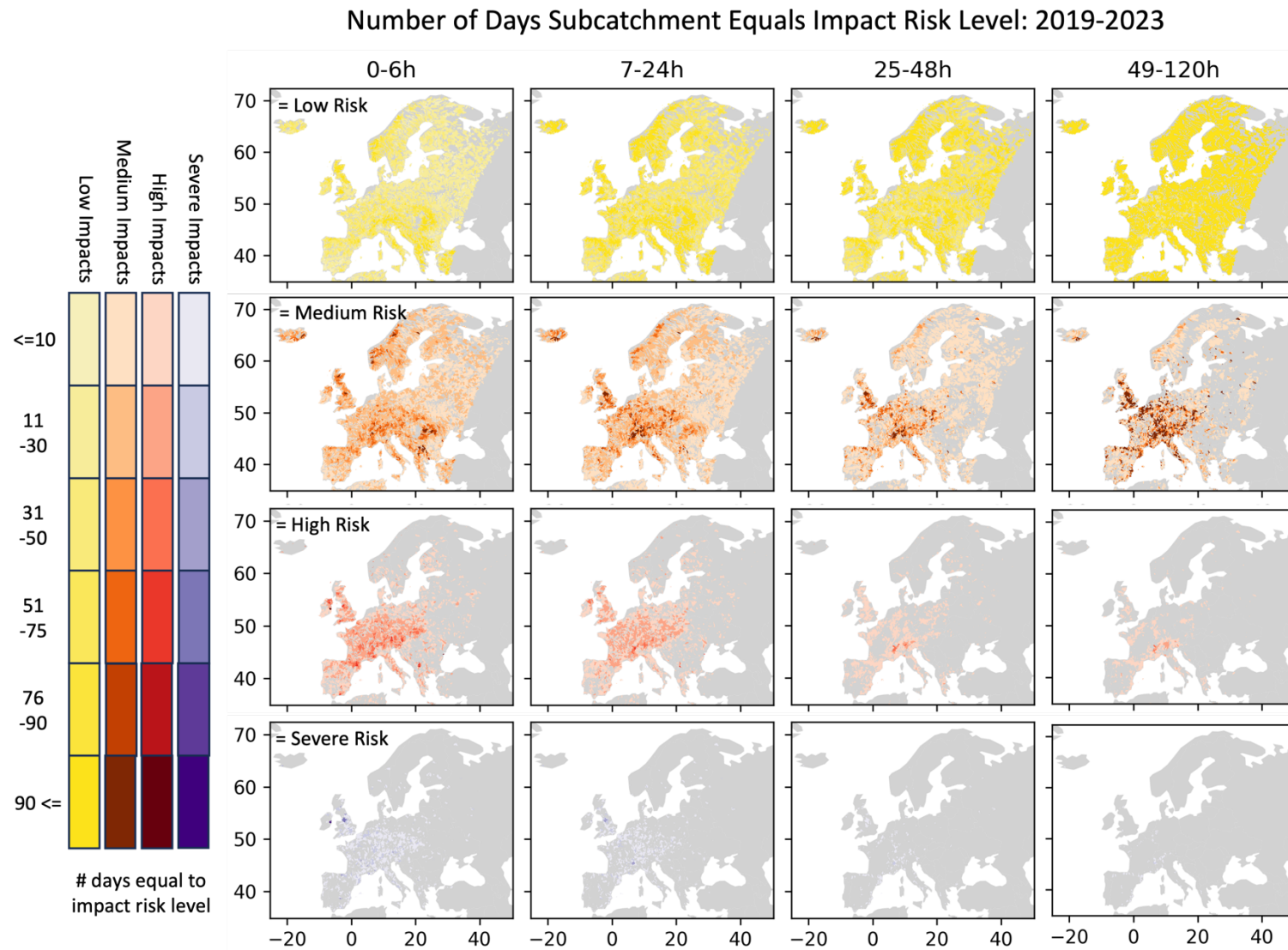


Figure 6 Number of forecast production days when each subcatchment had a forecasted impact level in each of the four impact categories.

The number of forecast production days between 2019-2023 when each subcatchment was highlighted in one of the four impact classes (yellow, orange, red and purple) was also analysed. Results showed subcatchments were highlighted most frequently as the yellow impact class, with purple impacts having the lowest frequency (Figure 6). This is expected due to the lower forecast hazard probability required to trigger the lowest impact levels. At the longest lead times, the frequency of subcatchments highlighted as yellow or orange impacts appears greater than at the shorter lead times. This is due to the greater forecast spread at longer lead times which reduces the forecast hazard probability causing a lower impact classification. At shorter lead times there may be a higher forecast hazard probability which results in a greater number of higher impact levels (red and purple). The spatial distribution of yellow impacts at longer lead times is more evenly distributed than orange impacts, this is likely due to spatial distribution of the exposure data, as orange impacts can only be triggered in areas with higher population density, as demonstrated by the concentration of orange impacts in western Germany and southern UK. Red and purple impacts are most likely to occur during the first two lead time ranges (0-6h and 7-24h), as mentioned above, because forecast uncertainty at longer lead times is too high to trigger these higher impact categories.

4.1.2 Historical Analysis of Flash Flood Notifications in Finland

The criteria provided by SYKE to receive notifications of flash floods from the EDERA forecast during the demonstration phase were:

- Lead time range: 7-48 hours
- Area of interest: all of Finland
- Minimum impact level: 2 (orange)
- Do not repeat a notification in a catchment, if the above criteria were already met within the previous 24 hours

These criteria were applied to the EDERA flash flood forecasts between 2019-2023 to identify the number of notifications that would have been issued to SYKE over the period. The findings could then be communicated to the stakeholder to see if this number was too many or too few and if modifications to the notification criteria should be made.

A total of 345 notifications would have been issued to SYKE over the period 2019-2023. Most of these notifications were generated between May to October in each year (Figure 7), which coincides with the period associated with flash flooding associated with heavy rainfall in Finland. However, according to flash flood events reported in the media and from SYKE, there were no reported events during 2019, 2021 and 2022, therefore it is likely that these notifications would be considered as false alarms.

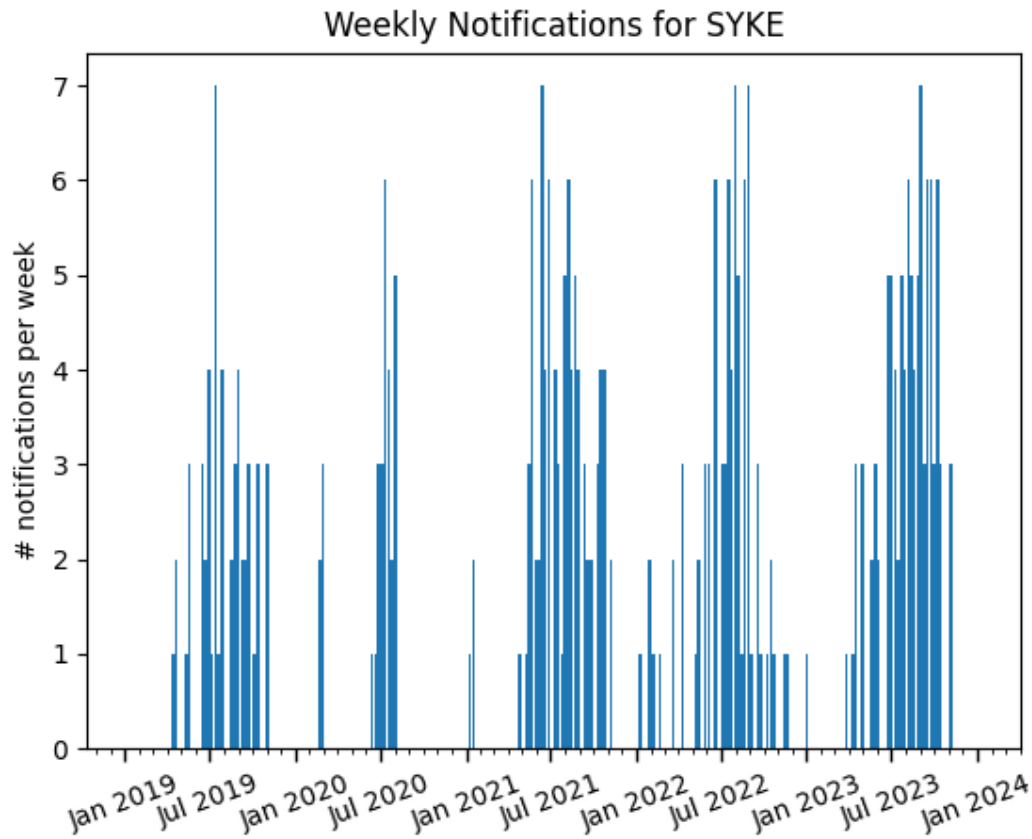


Figure 7 Number of notifications per week which would have been disseminated to SYKE according to their notification criteria.

The total number of notifications generated in each Finnish catchment, during the 2019-2023 period was also analysed. Results show that most notifications occurred in the larger basins in the south-east of the country and the Kemijoki catchment in the north (Figure 8). This is potentially an artefact of these catchments being large meaning there are more EDERA subcatchments within their boundaries. It is possible that if the catchments were evenly sized, the spatial distribution of notifications across the country would appear more even.

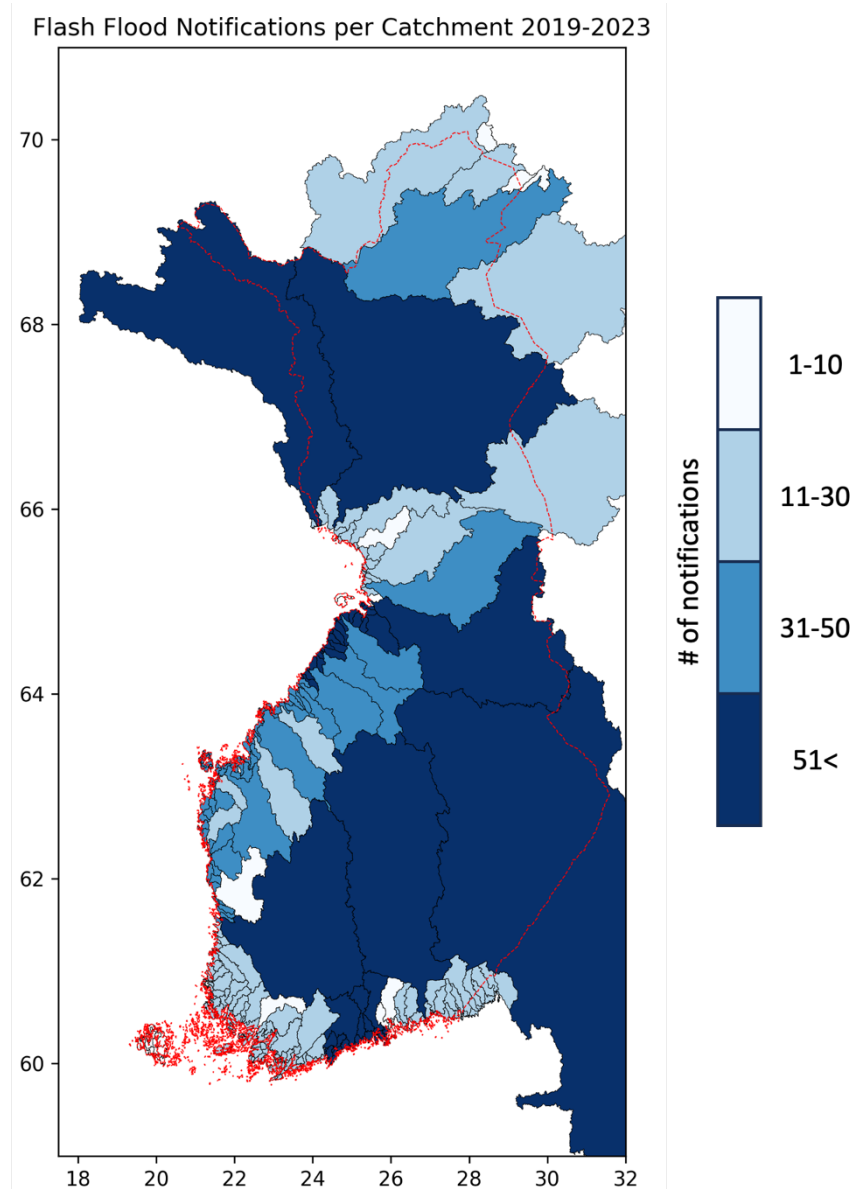


Figure 8 Total number of forecast production days when conditions in each basin met the criteria to generate a notification for SYKE.

The seasonal distribution of catchments which met the notification criteria was also analysed by repeating the above analysis but separately for each season. Results show the greatest concentration of notifications occurred during the summer and autumn, fewer notifications were generated in spring, only a few were generated in winter in the south-west (Figure 9). The larger catchments show a greater number of notifications during summer and autumn, but this is likely because their larger size increases the likelihood of one of the EDERA subcatchments within their boundary satisfying the notification criteria. The occurrence of some notifications in the south-west during the winter is unexpected as usually precipitation would fall as snow during this season, hence not triggering flooding. One possibility, as mentioned in section 4.1.1, is that at lead times beyond six hours the snow masking does not use the forecast snow probability product, which may have resulted in a failure to mask out these snowfall events. Therefore, future work to improve snowfall masking at longer lead times could prevent notifications which are likely to be false alarms during the winter.

Number of Seasonal Notifications per Catchment 2019-2023

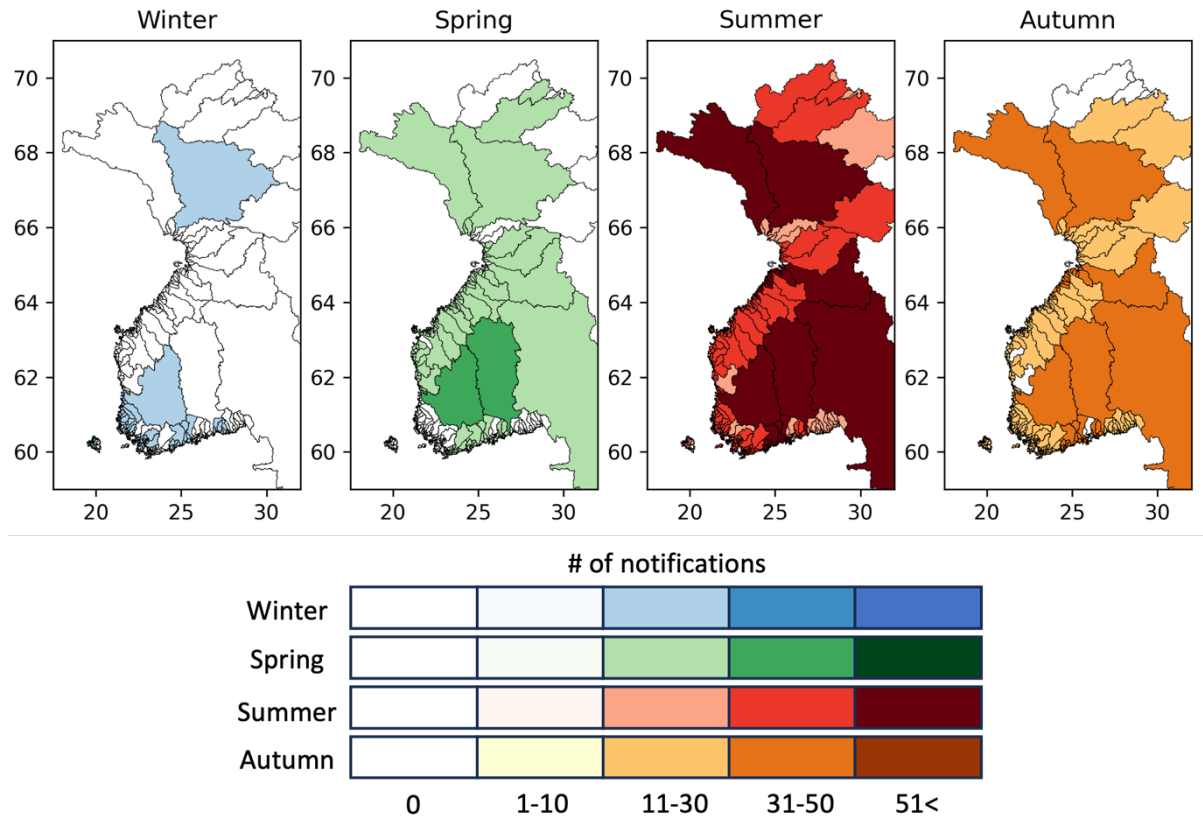


Figure 9 Seasonal distribution of the number of forecast production days between 2019-2023 when the forecast in each basin satisfied the notification criteria provided by SYKE.

4.1.3 Historical Analysis of Flash Flood Notifications in Andalusia

The AMAYA project partner provided criteria for receiving notifications separately for two different lead time ranges:

- Lead time range 0-6 hours:
 - Area of interest: Andalusia
 - Minimum impact level: 3 (red)
 - Spatial unit of interest: Andalusian counties
 - Do not repeat a notification in a county, if the above criteria were already met within the previous 24 hours
- Lead time range 7-48 hours:
 - Area of interest: Andalusia
 - Minimum impact level: 3 (red)
 - Spatial unit of interest: NUTS 3 administration regions
 - Do not repeat a notification in a NUTS 3 region, if the above criteria were already met within the previous 24 hours

Therefore, the analysis conducted for Finland in section 4.1.2 was repeated for the AMAYA partner to analyse the number of notifications which would have been issued between 2019-2023 for both lead time ranges above. At 0-6 hours AMAYA specified receiving notifications at the county level, initial investigations found the size of the counties to be too small compared to the size of the EDERA subcatchments because a notification in one subcatchment would highlight multiple counties, whereas AMAYA would prefer to know which of the highlighted counties is at risk. Therefore, this analysis was performed at the same scale of the NUTS 3 regions as the 7-48 hour notifications. In future work, the notifications could be generated at county level directly from the gridded 1 arcmin forecasts.

Over the five year evaluation period, 63 notifications were generated using the criteria at 0-6 hours lead time, whilst 96 notifications were generated using the 7-48 hours lead time criteria. The temporal distribution of the notifications was similar for both lead time ranges, both generated many notifications in the second half of 2022 until May 2023, whereas fewer notifications were generated during the first three years of the analysis (Figure 10).

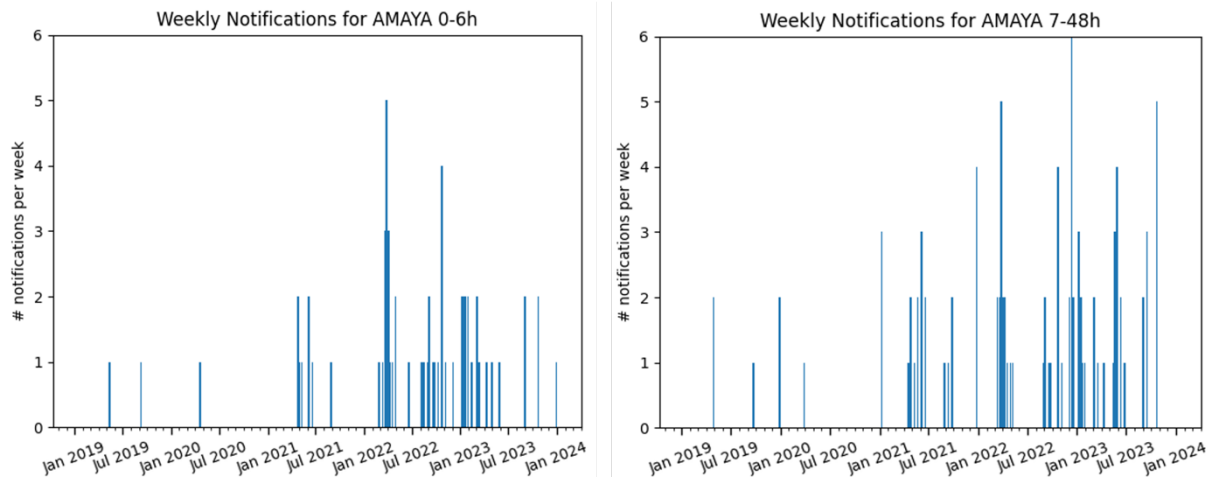


Figure 10 Number of notifications per week which would have been disseminated to AMAYA according to their notification criteria for 0-6 hours lead time (left) and 7-24 hours lead time (right).

The spatial distribution of notifications generated in each NUTS 3 region within Andalusia for each lead time range was analysed. At 0-6 hours lead time, the central part of the region has fewer notifications than the east, but at 7-48 hours lead time one of the central regions has one of the highest number of notifications (Figure 11). It would be expected that spatial distribution of notifications would be similar across the two lead time ranges, the different impact level thresholds should only affect the overall number of notifications rather than the spatial distribution. At 0-6 hours lead time, the forecasts are driven by a blend of radar and NWP precipitation forecasts, whereas the 7-48 hour lead time forecasts are forced only by NWP precipitation forecasts. The spatial differences in this analysis could suggest spatial inconsistencies in the rainfall distributions forecasted by the blended radar and NWP at shorter lead times compared to the NWP only at longer lead times. This issue merits further investigation in future work.

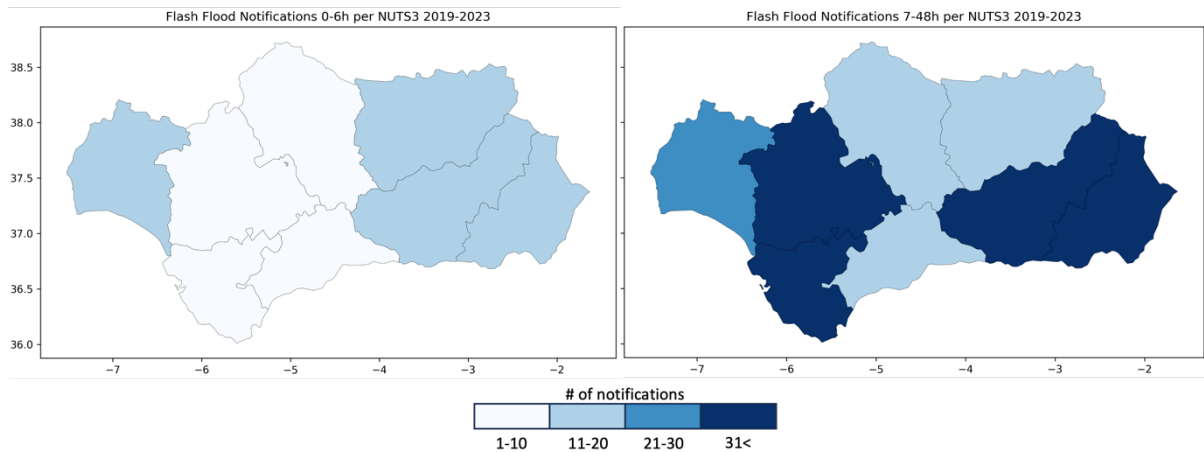


Figure 11 Total number of forecast production days when conditions in each basin met the criteria to generate a notification for AMAYA for 0-6 hours lead time (left) and 7-24 hours lead time (right).

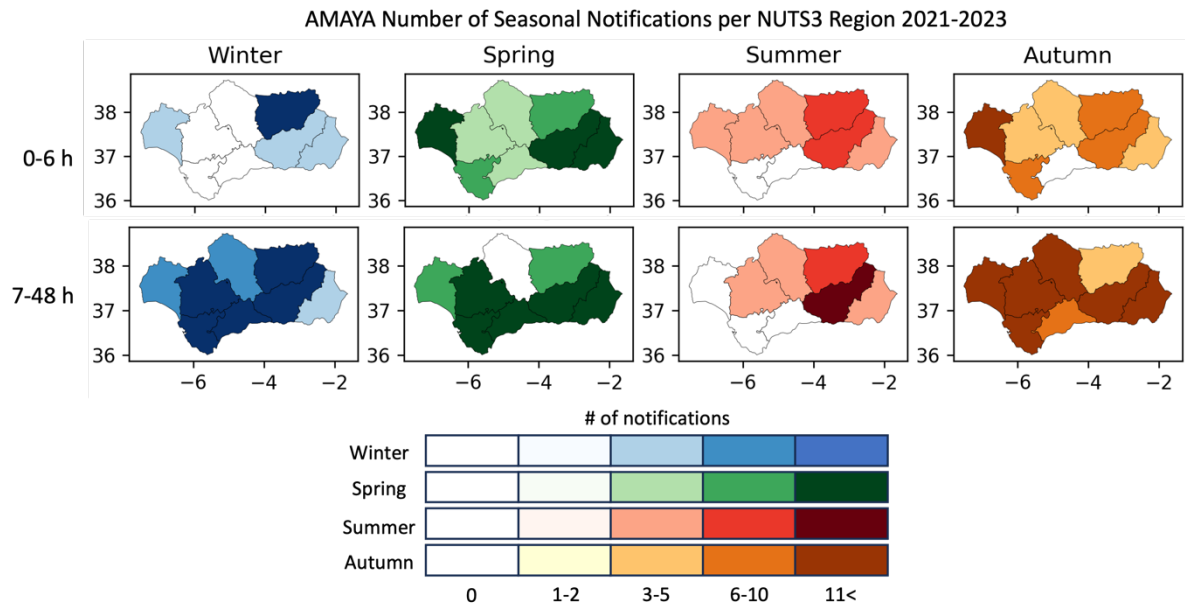


Figure 12 Seasonal distribution of the number of forecast production days between 2019-2023 when the forecast in each basin satisfied the notification criteria provided by AMAYA for 0-6 hours lead time (top) and 7-48 hours lead time (bottom).

The spatial distribution of notifications was broken down into the number of notifications generated in each NUTS 3 region in each season. As per the previous analysis, there were differences in the spatial distribution between the two lead time ranges, although during the summer both lead time ranges showed most notifications being generated in the east of Andalusia (Figure 12). At 0-6 hours lead time, the central regions had fewest notifications during winter and summer, whereas they received a few notifications in these regions during spring and autumn, the eastern part of the province received many notifications during all the seasons. At 7-48 hours lead time, during winter and spring most notifications were generated along the southern and eastern coastal regions, in summer most notifications were in the north and east whilst in autumn notifications were spread over most of the province. It is unclear if the spatial patterns described in either of the lead time ranges matches the seasonal distribution of rainfall across the province.

4.2 Evaluation against Media Reports of Flash Flood Impacts

In this section, the forecast products generated between 2019-2023 are evaluated against media reports of flash flood impacts using the methodology described in section 3 in each of the study regions of the project (Finland and Spain/Portugal).

4.2.1 Finland

Between 2019-2023 only three unique meteorological events associated with flash flooding impacts were reported in Finland, these occurred between in February 2020 and September – October 2023. It was not possible to robustly compute the recall, precision and f(beta) scores from only three unique meteorological events, therefore the number of hits, misses and false alarms from different impact level thresholds at each of the four lead time ranges was analysed. The results show at 49-120 hours lead time zero catchments exceeded any of the impact level thresholds, at 25-48 hours lead time impact level one was forecasted in some of the catchments but at the expense of several false alarms, impact level two was forecasted just once (Figure 13). At 7-24 hours lead time, applying an impact level threshold of three reduces false alarms with a small reduction in hits. At 0-6 hours lead time an impact level threshold of three has the same effect of reducing false alarms with only a slight reduction in hits.

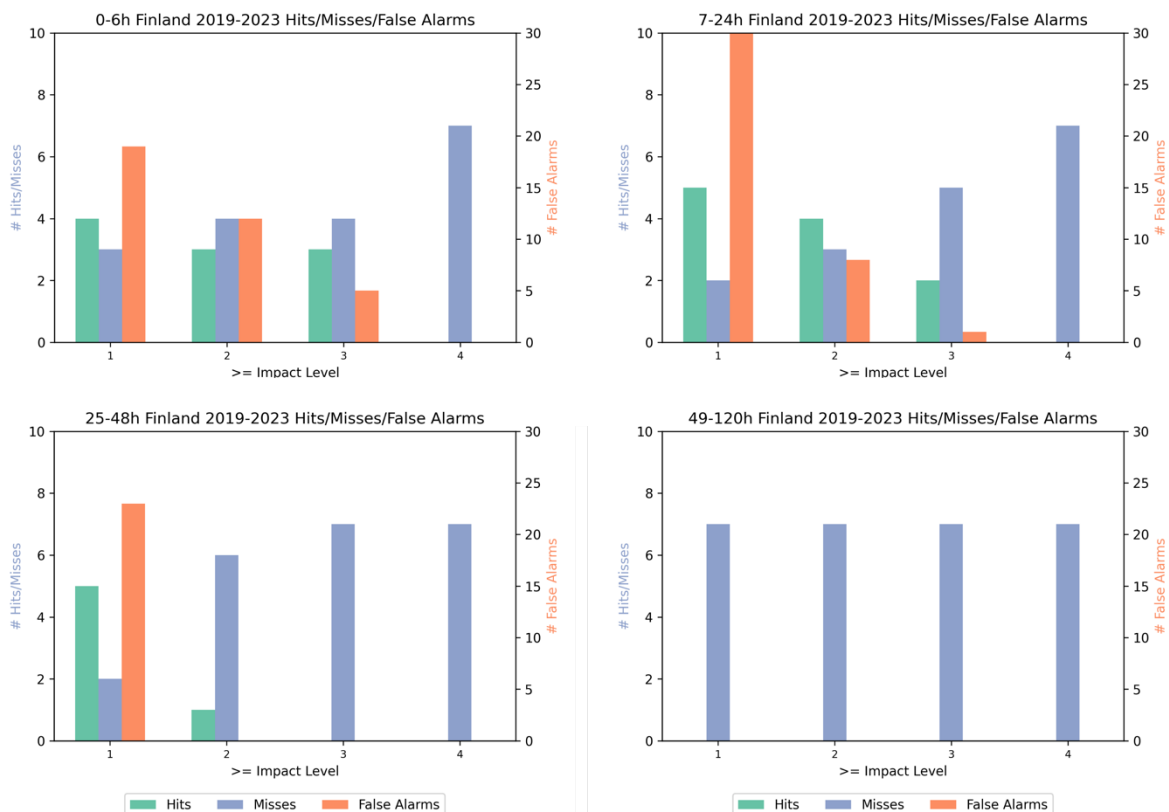


Figure 13 Hits, misses and false alarms from the evaluation of the EDERA forecasts applying different impact level thresholds against flash flood observations in Finland between 2019-2023 at different lead time ranges.

The SYKE stakeholder expressed an interest in receiving notifications from the EDERA forecasts between lead times of 7-48 hours. However, the analysis shows a mixed forecast performance

over this lead time range. At 7-24 hours lead time, good results were obtained by applying an impact level threshold of three which gave a good trade-off between hits and false alarms. However, at 25-48 hours lead time there was very little skill in the forecast as hits were only achieved at the expense of a greater number of false alarms. However, robust conclusions about the optimal notification criteria for the SYKE stakeholder require further analysis over a greater number of events.

The spatial distribution of the hits, misses and false alarms were analysed for the lead time range 7-24 hours and an impact level threshold of three. The results show that all three components are concentrated in the north east of the country (Figure 14), which is the same area where both the 2023 flood events occurred. The misses in the south of the country coincide with the location of the 2020 event. The plotting appears to show zero false alarms, however there is one small coastal catchment in the north east where a false alarm occurs, the catchment is too small to appear clearly on the figure.

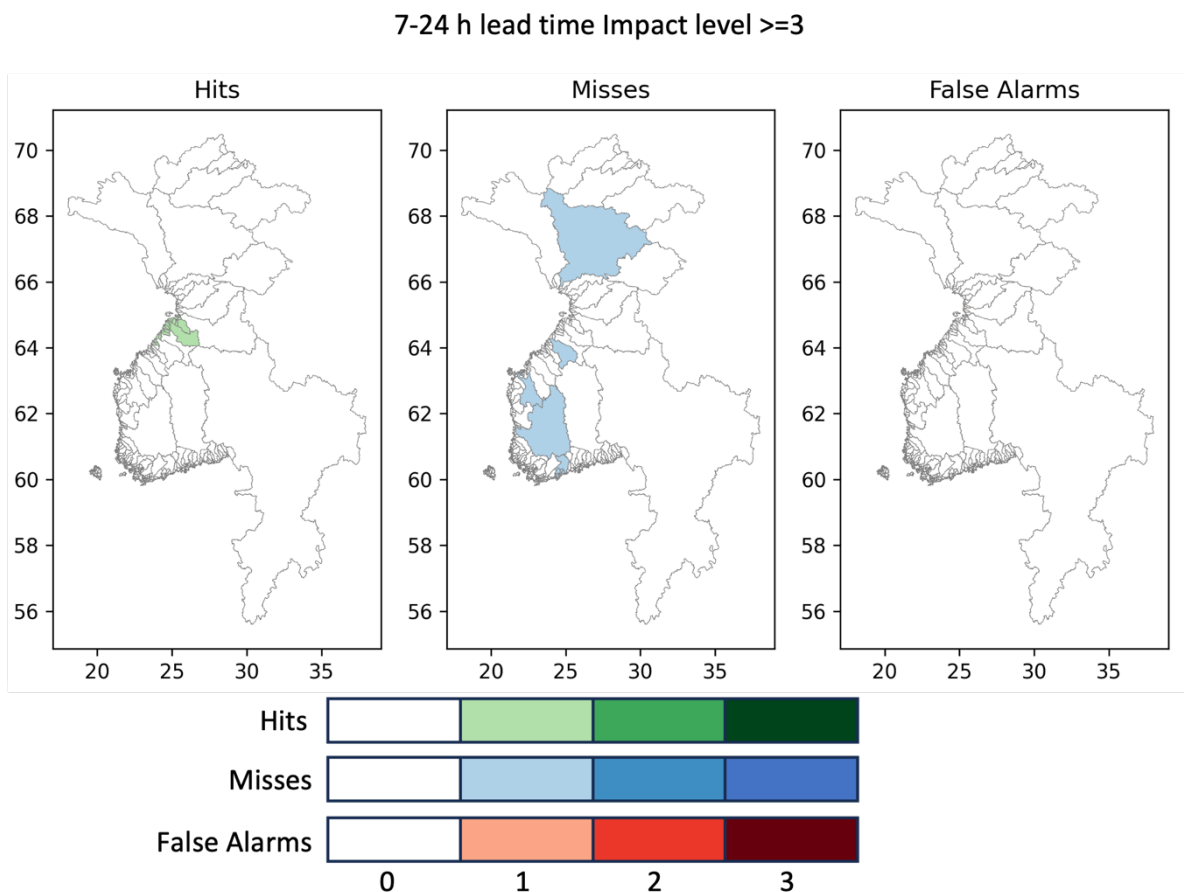


Figure 14 Catchment distribution of hits, misses and false alarms between 2019-2023 in Finland for 7-24 hours lead time applying a minimum impact level threshold of 3.

4.2.2 Spain and Portugal

Across the 38 unique observed meteorological events in Spain and Portugal (section 2.1.1), the recall was greatest at shorter lead times when using lower impact level thresholds (Figure 15). This is expected because applying a lower impact threshold will highlight more regions resulting in fewer misses. At the shortest lead time (0-6 hours), increasing the threshold from impact level 1 to

impact level 3, as requested by the AMAYA partner, reduces the recall from ~0.73 to ~0.50 (where 1.0 is the ideal value). However, applying the same impact level 3 threshold at longer lead times results in a greater reduction in the recall score, due to the increased number of missed events.

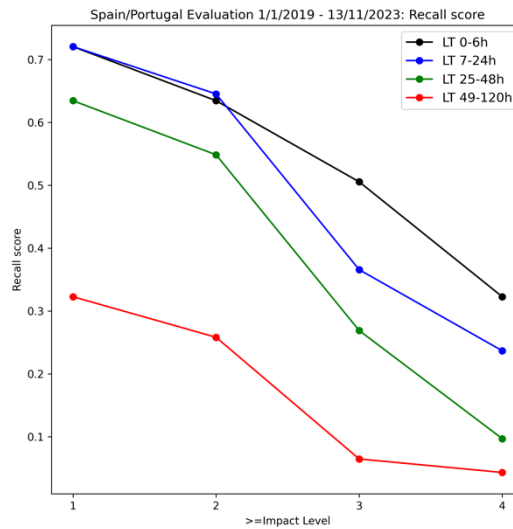


Figure 15 Recall score (hits / (hits + misses)) across Spain and Portugal between 2019-2023 for different lead time ranges when applying different minimum impact level thresholds.

Results from the recall score also needs to be considered against the forecast precision, a high recall value could mask an imprecise forecast which generates too many warnings. The precision scores over Spain and Portugal show that lower impact level thresholds result in lower values (Figure 16), whereas the inverse is true for higher impact thresholds. Therefore, lower impact values lead to the forecasts over-warning potentially affected areas. The greatest precision score was for the lead time range 49-120 hours at the impact level values of 3 and 4, the high precision values are due to there being zero or close to zero false alarms. However, this gives a false impression of skill as these high impact thresholds at this long lead time result in a very large number of missed events (Figure 18).

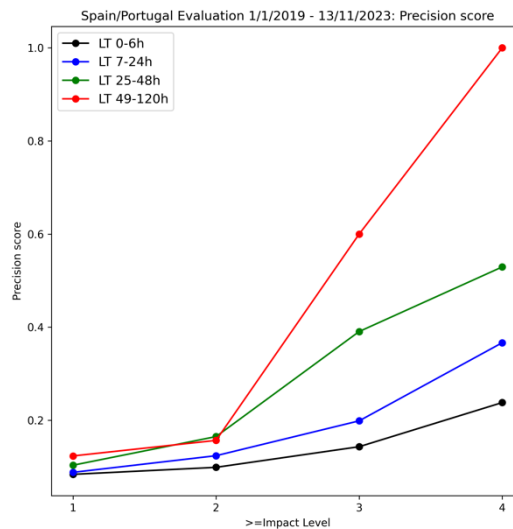


Figure 16 Precision score (hits / (hits + false alarms)) across Spain and Portugal between 2019-2023 for different lead time ranges when applying different minimum impact level thresholds.

Calculating the $f(\beta)$ score allows for the recall and precision scores to be combined to analyse the trade off between maximising the number of hits whilst minimising the number of false alarms. Over Spain and Portugal, the $f(\beta)$ scores are low across all lead times and impact levels thresholds (Figure 17), most likely a result of the large number of false alarms highlighted by the low precision scores. The highest $f(\beta)$ score was obtained at 25-48 hours lead time with an impact level threshold of 3. At shorter lead times from 0-24 hours the $f(\beta)$ score improved with increasing impact level, this could be caused by the increasing removal of false alarms at higher impact levels. At 25-48 hours lead time the $f(\beta)$ score dropped at impact level 4 possibly due to the large removal of the number of hits. At 49-120 hours lead time the $f(\beta)$ scores were low for all impact levels and decreased with increasing impact level, possibly due to the increased number of misses. Overall, the findings from the $f(\beta)$ scores would suggest that the skill is greater at medium lead times (7-48 hours) than shorter lead times (0-6 hours), this is surprising and contradicts the expectation that the shorter lead time should have higher skill due to its use of the radar rainfall information.

To investigate further, the number of hits, misses and false alarms obtained when applying different impact level thresholds for all lead time ranges were analysed. Results show that at all lead time ranges, applying lower impact level thresholds results in a higher number of hits but also a higher number of false alarms (Figure 18). The large number of false alarms would likely not be acceptable to a user as they greatly outnumber the number of hits. For both lead time ranges, increasing the impact level threshold reduces the number of hits but has a greater impact upon reducing the number of false alarms. At 0-6 hours lead time, increasing the impact level threshold to three reduces the number of false alarms by two thirds but reduces the number of hits by one third and the number of hits is larger than the number of misses. For 7-24 and 25-48 hours lead time range, applying a minimum impact level threshold of two reduces false alarms by nearly half, the number of hits is also reduced but still exceeds the number of misses. For 49-120 hours lead time range, only the lowest two impact levels generate any hits, the number of which are lower than the number of misses.

The above analysis shows that using the $f(\text{beta})$ skill metric is not necessarily useful to identify the optimal criteria for issuing notifications to users. This is possibly due to the large imbalance between false alarms, hits and misses, where large numbers of false alarms skew the importance assigned to the precision component of the $f(\text{beta})$ formula. The beta parameter can be modified to give different weighting to false alarms, however a sensitivity analysis with different beta parameter values showed the resulting $f(\text{beta})$ scores to be insensitive, most likely due to the overwhelming number of false alarms. Instead analysing raw values of hits, misses and false alarms from different impact level thresholds gives more insight into the trade-off between hits and false alarms. Therefore, it would be advised to users in Spain and Portugal that for lead times between 0-6 hours an optimal trade-off between hits and false alarms would be achieved using an impact level threshold of 3 and above. Whilst for longer lead times at 25-48 hours optimal performance can be achieved with an impact level threshold of 2 and above.

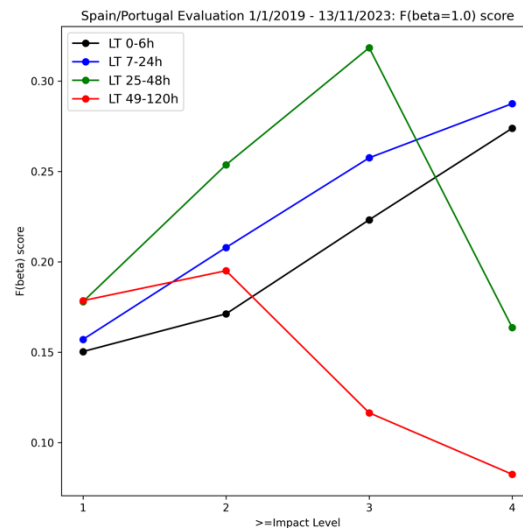


Figure 17 $F(\text{beta})$ score across Spain and Portugal between 2019-2023 for different lead time ranges when applying different minimum impact level thresholds.

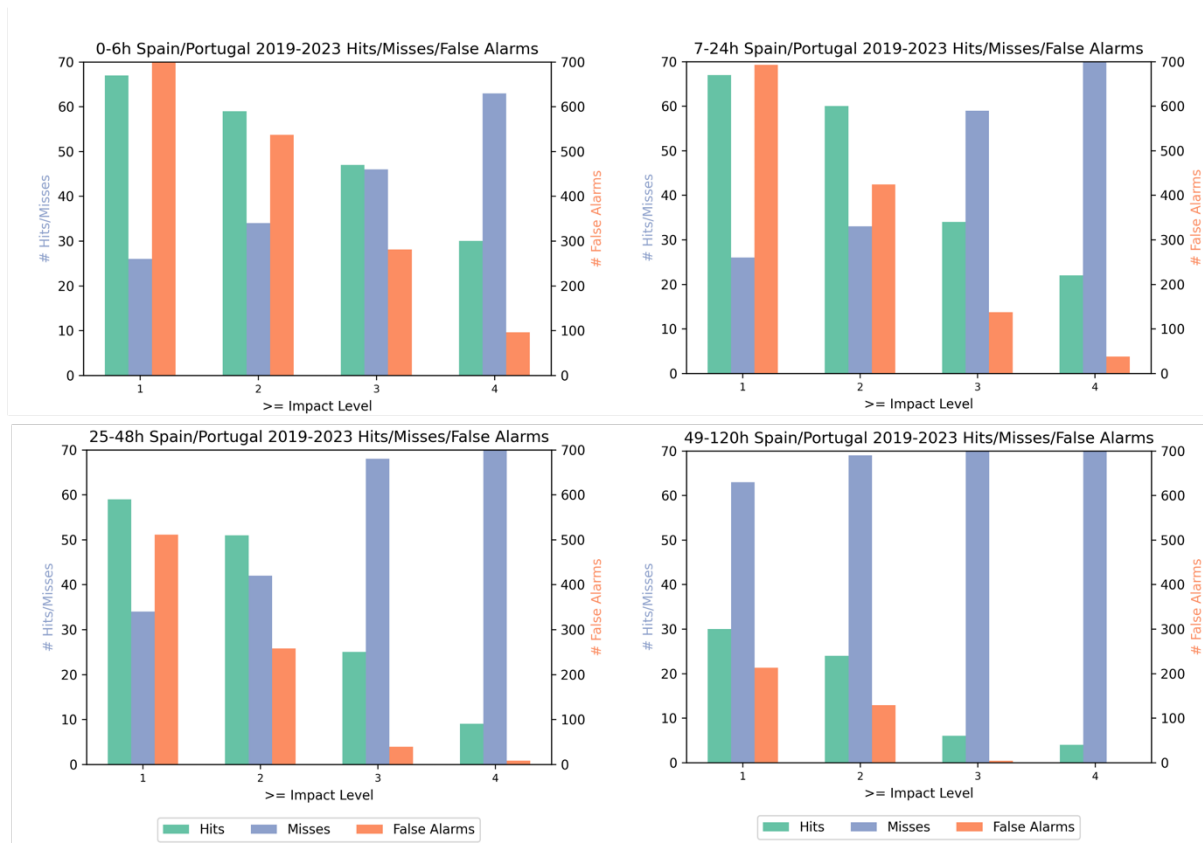


Figure 18 Hits, misses and false alarms over Spain and Portugal between 2019-2023 at 0-6 hours (top left), 7-24 hours (top right), 25-48 hours (bottom left) and 49-120 hours (bottom right) lead time ranges when applying different impact level thresholds.

A final analysis investigated the spatial distribution of hits, misses and false alarms in each NUTS 3 region at 0-6 hours lead time when applying the impact level threshold of three and at 7-24 and 25-48 hours when applying an impact level threshold of two. Data for 49-120 hours was not plotted due to the low number of forecasted events at all impact level thresholds. Results show that most hits occur in the north west, north east and central parts of the peninsula, false alarms are concentrated in the northern areas and misses are distributed across the region (Figure 19). There were a greater number of false alarms at 7-24 hours lead time across the peninsula, particularly the north east. In Andalusia, where the AMAYA project partner is located, fewer false alarms were recorded, however it is not possible to make any definitive conclusions about the forecast performance in this area relative to the performance across the whole peninsula.

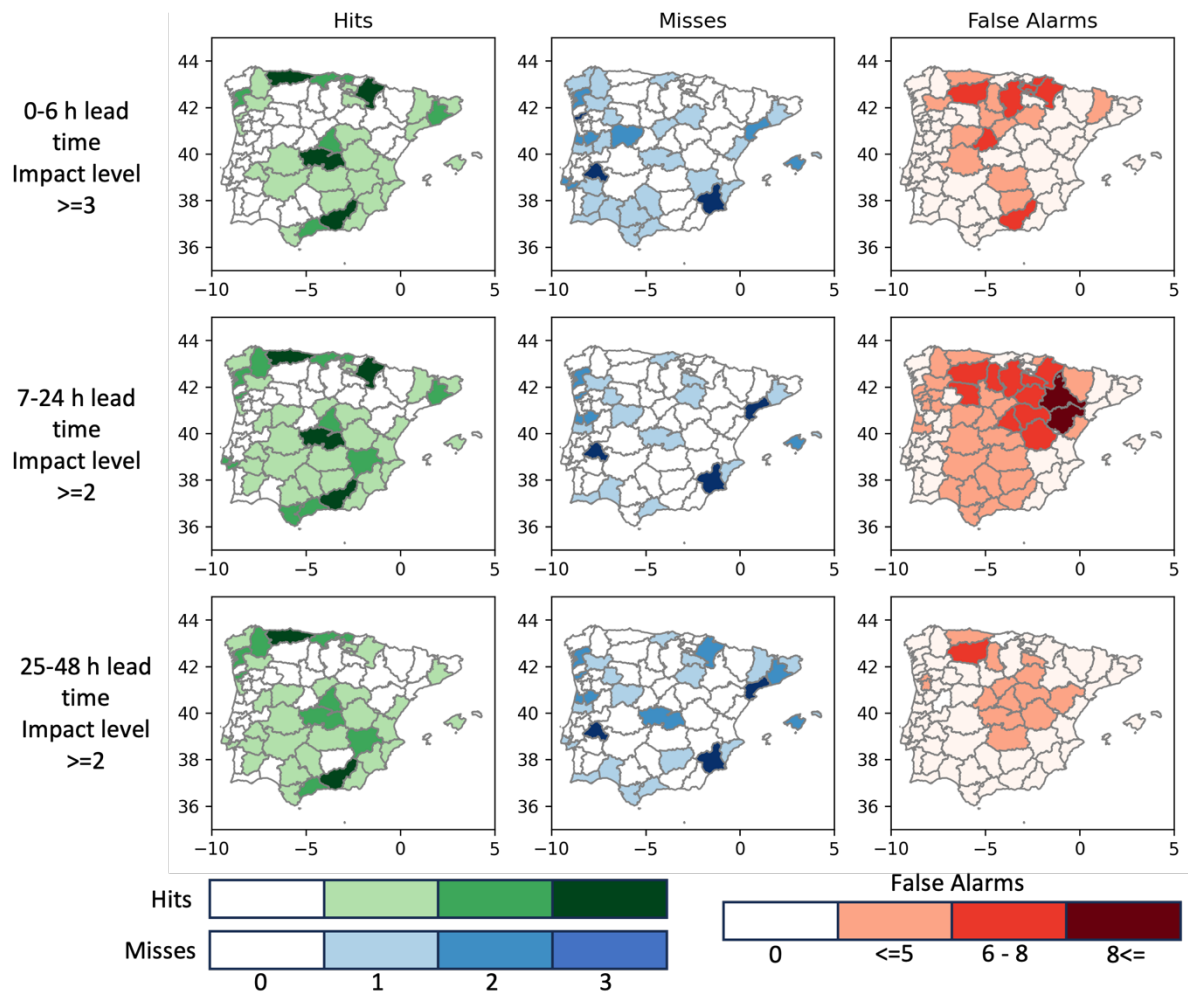


Figure 19 NUTS 3 regional distribution of hits, misses and false alarms between 2019-2023 in Spain and Portugal for 0-6 hours lead time applying a minimum impact level threshold of 3 (top) and 7-24 hours and 25-48 hours lead time applying a minimum impact level threshold of 2 (middle and bottom respectively).

5 Discussion

The pan-European diagnostic analysis of EDERA flash flood forecasts over 2019-2023 highlighted a potential issue where too many subcatchments during winter in northern areas were forecasted as being at risk at medium range lead time. Minimal risk would be expected in these areas as precipitation would mostly fall as snow. The masking of snowfall at medium range lead times may be the cause of this issue, currently a simple temperature threshold is applied to differentiate snow and rain but a different approach may be required. At shorter lead times the snow masking uses an NWP snow probability forecast produced by FMI, however this is not used at longer lead times. Future work should investigate if a snow probability product at longer lead times can be used to mask snowfall.

Both the diagnostic and evaluation analyses showed poor performance of the EDERA flash flood forecasts at medium range lead times greater than 48 hours. In Finland, the evaluation found that no observed event was correctly predicted beyond 48 hours lead time, between 25-48 hours only one event was correctly predicted but at the cost of many times more false alarms (Figure 13). In Spain and Portugal, the performance between 24-48 hours was slightly better, but beyond 48 hours the recall score was very low due to many missed events (Figure 15). At the pan-European scale,

the diagnostic analysis showed that beyond 24 hours lead time flash flood impacts were mostly predicted at one of the two lower impact levels (Figure 6). Therefore, if a stakeholder wished to receive notifications for forecast lead times beyond 24 hours they would need to carefully consider the balance between false alarms and misses when selecting the minimum impact level threshold criterion.

The increased spread in the position and magnitude of extreme precipitation in the ensemble NWP precipitation forecasts at these lead times may explain the limited performance at longer lead times. For example, the forecast spatial uncertainty can result in the extreme precipitation being located over different subcatchments. This means that over these subcatchments there could be a high likelihood of extreme precipitation but dividing it between different subcatchments results in a lower probability when calculating the flash flood risk separately in each subcatchment. Therefore, in future work calculating the probability of flash flooding at medium range lead times should be calculated across spatial scales larger than the subcatchments which are currently used.

The evaluation of the forecasts against observations focussed on results obtained when applying different lead time ranges and impact level values. Both can be configured by stakeholders when specifying criteria to receive notifications from EDERA forecasts. However, future work should consider evaluating additional criteria. For example, the forecasted probability of flash flooding was not considered in this study as it is combined with exposure data to estimate the impact level, which was evaluated. A stakeholder however may wish to base the notification criteria on the probability and then combine the forecast data with their own exposure information which may differ from the pan-European data used within EDERA. Another criterion which could be evaluated is the estimated affected population as this may align closer to the decision making workflow of stakeholders during a response to an emergency events. Therefore, future work should investigate offering more criteria to stakeholders when configuring the notifications they receive from the forecasts.

6 Conclusions

The analysis of EDERA river flash flood forecasts between 2019-2023 in the Finland and Spain/Portugal study sites has shown a good ability of the forecasts to represent observed events at lead times less than 24 hours ahead. Between 24-48 hours lead time the performance was variable, good results were obtained in the Spain/Portugal site but results were poor in Finland. Beyond 48 hours lead time, the forecasts were only able to predict flash flooding with a lower impact level due to the increased forecast uncertainty which could lead to many false alarms. Therefore, it would not be recommended to issue notifications from the EDERA forecasts beyond lead times of 48 hours.

For the AMAYA and SYKE stakeholders, the results from this study can provide them with the following guidelines to establish their criteria to receive notifications:

- For SYKE, the maximum forecast lead time should be reduced from 48 to 24 hours, the minimum impact level threshold should be increased from 2 to 3

- For AMAYA, the original notification criteria they suggested at both lead time ranges (0-6 and 7-48 hours) gave good results. At 0-6 hours it is recommended that notifications are also issued at the NUTS 3 spatial scale rather than the county level.

Future studies should investigate improving the performance of the forecasts at lead times beyond 48 hours by considering different methods to calculate the flash flood probability which considers the increased spatial uncertainty of the forecast. Offering additional criteria for stakeholders to configure their notifications could also improve their performance and may also help with their integration into stakeholders' decision making workflows during emergency events.

References

- [Error! Reference source not found.] Saltikoff, E., Haase, G., Delobbe, L., Gaussiat, N., Martet, M., Idziorek, D., Leijnse, H., Novák, P., Lukach, M., & Stephan, K. 2019. OPERA the Radar Project. *Atmosphere* 2019, Vol. 10, Page 320, 10(6), 320. <https://doi.org/10.3390/ATMOS10060320>
- [2] Park, S., M. Berenguer, and D. Sempere-Torres, 2019: Long-term analysis of gauge-adjusted radar rainfall accumulations at European scale. **Journal of Hydrology**, 573, 768–777.