## On the importance to examine ROC curves and not only AURC when verifying rainfall forecasts against non-rainfall observations

## On the discrimination ability of ENS and ecPoint forecasts between flash flood events and non-events

In Section 5.1, it was highlighted that the AURC values for ENS at the 85th percentile are similar to the AURC values for ecPoint at the 99th percentile. Therefore, the ability of ecPoint to discriminate between flash flood events and non-events when considering high rainfall events is similar to the discrimination ability of ENS when considering high rainfall events.This means that a forecaster could get a sense about whether there is going to be a flash flood somewhere in “La Sierra” by using high probabilities high probabilities for smaller rainfall events using ENS, or using small probabilities of bigger rainfall totals from ecPoint. Is arguable however, that the second situation would be more amenable to forecasters. Furthermore, if the forecaster references the ecPoint-Rainfall forecasts to actual rainfall values, they are likely to appear more representative of the localized extremes. This would be consistent with (Hewson & Pillosu, 2021) findings, which showed that ecPoint-Rainfall forecasts provide a more reliable and skilful representation of point-based rainfall observations than ENS, including extreme rainfall (>= 50 mm/12h).

We pitch two ROC curves in which the AURC for ENS at 85th percentile is similar to the AURC for ecPoint at 99th percentile. The two ROC curves are different. In terms of what the different probabilities are contributing to the ROC area, the points that correspond to ecPoint’ low probability values are those contributing to the ROC area. In the ENS’s ROC curve, the high probability values are those contributing mainly to the area under the ROC curve. The small probabilities, from 1% to 10%, are not contributing to the area under the ROC because they ley on a straight line (one could stop the ROC curve at the 10% probability point, draw a straight line, and the area would be the same).

## On the overlapping ROC curves for ecPoint and ENS in “La Costa”

(Bouallegue & Richardson, 2021) argue that if the ROC curves for a post-processed and raw forecasts overlap with the exception of only a couple of points, the post-processing might not be altering the underlying information content in the raw forecasting system. Instead, the improvement seen on few points in the ROC curve for the post-processed forecasts might be only attributed to a change in the frequency of the events (i.e., due to a bigger number of ensemble members, and therefore a bigger spread, the post-processed forecasts exceed the event-threshold more often than the raw forecasts). (Bouallegue & Richardson, 2021) arrive to this conclusion by analysing the ROC curves for raw and post-processed forecasts that add information on sub-grid variability to the raw forecasts and increase the number of raw ensemble members. They compared ROC curves and AURC calculated with the trapezium and the bi-normal technique. The ROC curves computed with the latter technique do not differ much (concluding that the post-processing technique does not alter the underlying discrimination ability of the raw forecasting system) while the ROC curves computed with the former technique overlap perfectly with the exception of few points (concluding that the ROC curve computed with the trapezium technique provides the false illusion of an improved underlying discrimination ability of the post-processed forecasts over the raw ones).

The ecPoint post-processing technique not only increases the number of ensemble members available to users (i.e., ecPoint provides a bigger spread than the raw forecasts). It also adds information on the possible rainfall sub-grid variability to produce forecasts for a point within the grid-box. If the information added by the post-processing technique regards also sub-grid variability (and not only, for example, bias correction), the raw forecasts will still not show a similar spread by simply increasing its ensemble members because, even a bigger ensemble, would still represent average forecasts over the grid-box which will not coincide with point-based rainfall observations unless the sub-grid variability is small. The two forecasting systems will still produce forecasts with different resolutions (i.e., grid-box scale and point-base scale) even when the spread of the raw ensemble is increased. This difference between the resolutions of ecPoint and raw NWP models has been shown in Pillosu et al. (2021) for rainfall case studies in Costa Rica and Hungary. It is important to notice that this distinction is valid for ecPoint and for whatever other post-processing technique that adds information on sub-grid variability. Therefore, it is valid also for the post-processing technique used by (Bouallegue & Richardson, 2021), and described in (Ben Bouallegue *et al.*, 2020).

Another important consideration regards the fact that the information provided by the bigger ensemble could be useful for a certain type of users, and they will not have access to that type of information if they use only the raw ensemble. (Bouallegue & Richardson, 2021) show that the potential economic value of the post-processed forecasts is higher than the one for the raw forecasts only for users with small cost-loss ratios (typically less than 2%) in case of overlapping ROC curves with the exception of few points closer to the top right corner of the diagram. This means that typically users who would incur very large losses when a rare event happens and not protection was taken would benefit of the post-processed forecasts (Richardson D., 2000). Even in the case that the resolution of the two forecasting systems (raw and post-processed) were similar and the ROC curves were proved to not be different with the exception of few points since the post-processed has a bigger number of members, when more points can be added to a ROC curve by having a greater number of ensemble members, the area will increase and so will the value for users (Richardson D., 2000). This is because the users will be actually able to use the extra information provided by the post-processed forecasting system that is not available in the raw ensemble. One could parametrize the ROC curves (for example using the bi-normal technique) to demonstrate the potential value that would be achieved if all possible probability thresholds could be used (i.e., an infinite ensemble of forecasts) to estimate the benefit that could be achieve by the raw forecast if it had a larger number of ensembles. Even in the case the two ROC curves (trapezium and bi-normal) would be the same, the reality stays that the raw forecast has a smaller number of ensembles that might not provide to users the full picture of what could happen (e.g., rainfall event exceeding 50 mm/12h) if the uncertainty of the event is high (typically for rare events). Furthermore, to have only one more point in the ROC curve like in the case of the raw ENS and ecPoint (that goes from the maximum 98th percentile for ENS to the 99th percentile for ecPoint), the number of ensemble members should double which comes to a high computational cost which might not be affordable. Here the focus is to show what are the practical benefits on using two forecasting systems with different configurations to forecast flash floods. The focus is not to compare their underlying discrimination abilities. The results in **Figure 8** show that the raw ENS does a good job in forecasting rainfall that might cause flash floods, and would probably fully overlap with the ROC curve for ecPoint if it had 100 members. However, increasing ENS to that number of ensemble members would come with a very high cost that cannot be afford with the current computational resources available.

## On the need to keep improving the collection of observational datasets to better verify new flash flood forecasting systems

There is a need  to differentiate flash flood from other type (to better understand how they are triggered). It is possible with the description and location of historical events from disaster datasets.

this study aims at providing a more organic verification analysis for flash floods in Ecuador than other studies that base their conclusions mainly on case studies (Raynaud *et al.*, 2015) or in rainfall verification (Park *et al.*, 2019). This is mainly due to the fact that there are not many available flash flood databases with good spatial coverage that can be used by researchers for forecasts validation (Gaume *et al.*, 2009; Kruczkiewicz *et al.*, 2021), while more detailed information might be available for single flash flood events and spatial/time coverage of rainfall observations is much better in quantity and quality, especially in Europe (Haiden & Duffy, 2016) and USA (Zhang *et al.*, 2011). While using case-study-based verification is sometimes the only possible verification methodology, it provides only a taste of how newly developed rainfall forecasts could be used in flash flood forecasting. This study verifies ecPoint-Rainfall forecasts in Ecuador thanks to the recent development of a comprehensive flash flood database carried out by Kruczkiewicz *et al.* (2021).

Flash floods constitute a significant risk in Ecuador due to the climatological and hydrological characteristics of the country, with the exception of the Amazon region where large-scale riverine floods are the dominant type (Kruczkiewicz *et al.*, 2021b). This is consistent with *flood* reports from international databases such as EM-DAT (CRED, 2019), Desinventar (UNDRR, 2021), and FloodList (Davies, 2019), which indicate that *floods* are not only the most frequent natural hazard in Ecuador, but they are also the most impactful. For example, the most severe flood events in 1982-1983,1997-1998, 2008, and 2012 caused economic losses estimated at over US$ 7.140 million (Galarza-Villamar *et al.*, 2018). Although specialized, these international databases are likely underestimating the real frequency of (flash) flood events as they tend to report mainly widespread, high-impact events. Ecuador’s Servicio Nacional de Gestión de Riesgos y Emergencias (SNGRE) reported that 2268 floods happened between 2014 and 2019, and 50% of the deaths in that period were due to those floods (SNGRE, 2019). EM-DAT provides only 4 flood reports for the same period, and only 1 is classified as flash flood. In addition, a web search in Spanish language shows that there are many more flash floods events that, due to being localized, with smaller impacts, do not reach international English-written news or databases. While (Pinos & Quesada-Román, 2022) show an increasing trend in the publication of flood-related peer-reviewed literature in Latin America that might continue in coming years, to the best authors’ knowledge there are no peer-reviewed statistics that describe the occurrence of flash floods in Ecuador.

Flood databases still do not contain enough data to clearly show the value of flash flood forecasting systems. Results of general verification can appear inconclusive. Therefore, such studies need to be complemented by relevant case studies to present the details that are lacking in the observational datasets.

In order to improve flash flood forecasts, it is also paramount to also improve the collection of flash flood observations. Currently they are mainly based on ground reports or radar data, which is available only in limited areas. Satellites has a huge potential since it will allow researchers to explore how flash flood risk may differ from one country, and city, to another (Andrew, blog post).

## On the need of climatologies for hydrological parameters to implement a similar system around the world without the need of local knowledge of thresholds

The Flash Flood Guidance with Global Coverage (Georgakakos *et al.*, 2021) is a successful story of the implementation of flash flood guidance in different countries. However, their implementation requires an involvement from the side of the country that might not be available because data for the implementation of the system or expertise might not be available.

This system offer the possibility to have a flash flood forecasting system with a continuous global domain that can provide flash flood forecasts around the world with no need of extra local data for the implementation in a region provided climatologies are available to define at least rainfall thresholds to identify “extreme” rainfall events that might lead to flash floods. If such climatologies would also include information on hydrological parameters that also modulate the occurrence of flash floods, it would also be possible to improve further the refinement of areas at flash flood risk.

The second benefit would also include the fact that non-local experts would be able to provide flash flood guidance for diverse regions. This is important in large-scale projects such as FbF or Aristotle. It would b impossible for a person to know the details that drive to flash floods in different regions of the world, but by having information on the climatology of rainfall or other hydrological parameters would allow that person to produce forecasts for those events in advance and help decision makers whether release resources to help the region at risk in help the event actually happens. (Georgakakos *et al.*, 2021) highlight how important would be to have a multidisciplinary team in meteorology, hydrology, remote sensing, and computer science to develop FFGS. The authors here claim that to develop a flash flood forecasting system with global coverage using ecPoint-Rainfall and climatologies for hydro-meteorological parameters would not require the presence of such expertise. However, the system would benefit from continuous feedback from users around the world which would lead to improvements in the forecasts as it happens for ECMWF IFS.

## On the use of the proposed system to foster humanitarian action in flash flood prone areas

Hydro-meteorologists claim that the exact place and time of the flash flood occurrence can be known only an hour before it transpires (Associated Programme on Flood Management, 2007).

This significantly reduces the possibilities for the early warning of crisis services and residents, particularly considering the fact that in many countries hydro-meteorological services have a limited range of measurements and forecasting tools at their disposal.

As a result, it is challenging for the concerned authorities and communities to respond timely and appropriately.

In an effort to minimize flood impact and the impact of all disasters, international bodies have come together to set goals and priorities for disaster risk reduction, as presented in the Sendai Framework for Disaster Risk Reduction (reference).

Despite the significant economic and human impacts of flooding, a portion of flood losses are widely considered avoidable through improved disaster risk reduction methods (Push, 2004). Government agencies and relief organizations have historically prioritized disaster relief, allocating the majority of financial resources to response efforts in a reactionary mode (Coughlan de Perez et al., 2016). However, empirical evidence demonstrates that shifting from a response-based approach to a more proactive operational paradigm with additional emphasis on preparedness has shown to save lives and reduce response costs (Braman, 2013; Perez, 2016; Aguirre et al., 2019; Braman et al., 2013; Golnaraghi, 2012; Gros et al., 2019). The continued improvement of short- to medium-range forecasts of natural hazards such as extreme rainfall and floods allowed the establishment of forecast-based early actions (FbA) initiatives, which are now recognized as a critical component of disaster risk reduction (IFRC, 2009). The goal of FbF is to trigger targeted action in the time between the issuance of a forecast and when the potential disaster occurs. The overarching goal is to prevent impacts and reduce human suffering. Since 2014, the German Government and Red Cross RedCrescent have been working on a new system called Forecast-based Financing (FbF), which aims at using hydro-meteorological forecasts to anticipate possible impacts in risk-prone areas and make resources for certain humanitarian actions automatically available before an event (Wilkinson et al., 2018). A key element of FbF is that the allocation of financial resources is agreed upon in advance, allowing for contingency plans to be designed and decisions to be outlined well before that stressful period just hours or days before a disaster occurs. Using historical records of extreme floods, we can analyse forecasts before those events to get an idea of how well they performed. Once we do that, we can develop a ‘trigger’, which is a collection of contingencies that need to be satisfied in order for donors to release funds and humanitarian organizations to take action.

While FbF has successfully financed already several interventions before severe riverine floods occur in different parts of the world (Jjemba *et al.*, 2018; Aguirre *et al.*, 2019a; IFRC, 2019; Emerton *et al.*, 2020), establishing a similar system for flash floods has been proved to be more complicated. The application of FbF systems have been limited predominantly as a result of moderate forecast performance and significant uncertainty in the case of flash flood forecasts. As forecasts improve, we’re finding that it is possible to get an idea of not only where a flash flood may occur, but where the impacts may be the highest. Based on this information, we can justify releasing funding to take preparedness actions. However, I should note that the lead time is much shorter to forecast flash floods compared to other flood types (Kruczkiewicz *et al.*, 2021b).

In addition, in case of lower probability, only less elaborate or intensive ‘low regret’ actions, such as refresher trainings, are taken. Sometimes early actions will be taken but the expected extreme weather event will not occur, so the action will be ‘in vain’. Therefore, the system will be designed so that more resource-intensive, elaborate, or disruptive actions, ‘high regret’ actions like evacuations, will only be taken when the probability of the extreme event is high. However, flash floods are typically always forecasts with only small probabilities. If FbF protocols do not adapt to this specificity of flash flood events, actions such as evacuation of people, might not be taken and life will continue to be lost. A late response can be two to six times more costly than actions in vain (Cabot Venton, 2012). Hence, while FbF funding is limited and there is not willingness to act in vain, over time, the negative consequences of not taking early action would be significantly greater than occasionally acting although the extreme event does not occur (reference). Also, Humanitarian assistance after disaster strikes is far more costly than investing in medium- to short-term anticipatory actions reducing impact and losses caused by disasters (reference).

In addition to short term weather forecasts to take immediate action, medium- to long-range climate forecasts have demonstrated their utility in improving preparedness protocols, resulting in reduce mortality, morbidity, and resource demands (Braman, 2013).

On the one hand, climate researchers and weather experts are able to determine the probability of extreme weather events for specific regions based on forecasts up to six months in advance (reference). Thus, many climate-related hazards, such as floods related to El Nino or La Nina, can be predicted up to sub-seasonal and seasonal scales. The question of when to initiate FbA requires integrating a hazard forecast with vulnerability and exposure information to estimate the impact of an extreme event. Weather-dependent risks can also be predicted with increasing accuracy. For example, early warning systems for slow-onset floods (riverine, for example) have significantly improved over the past 50 years. Thus, forecasts for this type of floods are increasingly used to help reduce the impacts of floods in vulnerable communities. See the example of GloFAS (Bischiniotis *et al.*, 2019). On the other hand, the scientific community struggles extending the predictability of other types of events beyond few days. This is the case for example of localized extreme rainfall which is typically the triggering event of severe flash flooding.

Forecast performance, uncertainty and hazard type dictate the range and extent of potential early actions available. While longer lead times allow for a greater range of potential early actions (Bazo, 2019), this must be balanced against corresponding increases in forecast uncertainty. One solution would be to make more use of probabilistic forecasts.

A key element of FbF is that the allocation of financial resources is agreed in advance, together with the specific forecast threshold that triggers the release of those resources for the implementation of early actions. One commonly used method to trigger early action is to define a forecast threshold above which impacts are likely to occur based on historical data. This method accounts for the probabilistic nature of forecasts by requiring a predetermined level of forecast confidence. Skill in detecting events is highly dependent on the threshold probability required to trigger early action. In general, a lower threshold for action will result in instances of worthy action but also more actions in vain. Conversely, a higher threshold for action will prevent false positives yet it will reduce the likelihood that early actions will be taken when needed.

When linking early action based on probabilistic forecasts to the occurrence of extreme events, four scenarios are possible (**Table 1**) where worthy action and worthy inaction are preferred. Therefore, forecast models that proficiently predict extreme events at lead times permitting early action are critical for minimizing false alarms and misses (**Table 2**). In particular, the risk of acting in vain, when early action is initiated but an extreme event fails to materialize (Lopez, 2017) is often viewed as a major barrier to scaling up FbA (Tanner, 2019). This tolerance for false alarms when implementing early action is an open question for decision makers and may depend on numerous technical, institutional, and political factors outside the scope of this study.

While FbA was initially applied to acute and slowly evolving threats like tropical cyclones, more recent efforts have targeted hydrological threats including extreme rainfall and flooding (Gros, 2019). However, while in the case of riverine floods, one can produce forecasts with relatively high probabilities (e.g., greater than 70 %), in the case of flash floods is almost impossible to provide forecasts with probabilities greater than 10%. This is not only due to deficiencies in the rainfall forecasts from global models. This is also due to the rare nature of the localized rainfall events that generally cause flash floods (Legg & Mylne, 2004; Goodwin & Wright, 2010; Hitchens *et al.*, 2013). Therefore, efforts to create a comparable system for flash floods has lagged behind (Kruczkiewicz *et al.*, 2021a).

Forecasts of rainfall or river discharge are an important tool for flood preparedness, providing information about the timing and magnitude of flood peaks. In contrast to deterministic forecasts, probabilistic forecasts provide a distribution of expected future rainfall totals or streamflow to account for uncertainties in future weather and hydrological model parametrization. While probabilistic forecasts can better represent uncertainty of future conditions, achieving widespread operational use of probabilistic river forecasts in the disaster risk management sector remains a challenge. This has been attributed to misuse of forecast information (Pielke, 1997) as well as a lack of training on how to translate probabilities into preventive actions and uncertainties regarding probable impacts (Hoss, 2016). From a risk communication perspective, the adoption of a new technology (i.e., probabilistic river forecasts) can be influenced by numerous factors including trust, ease of use, expected benefit and social pressures (Siegrist, 2000; Venkatesh, 2003).

An advantage of probabilistic forecasts over their deterministic counterparts is the ability to quantify uncertainty. Ernst (2018) found that forecasts providing additional details about upcoming events helped emergency managers better prioritize resources for preparedness. Longer lead time forecast have also been leveraged for preparedness. However, a literature search suggests that very few (Fundel et al., 2019; VanDyke 2020) focus on EM perceptions of probabilistic river forecasts specifically. However, few studies (Rayner 2005; Hoss 2014) show that suggest emergency managers are more likely to use observations (e.g., radar or river discharge) instead of model outputs in decision making because the uncertainty inherent in probabilistic forecasts are more difficult to interpret. As consequence, the lead time with which warnings are provided are reduced to few days to few hours in case of flash flooding.

## On the benefits of using ecPoint post-processing technique rather than other techniques

A remarkable drawback is the recurrent need to performing statistical reanalysis every time a component of the source large-scale NWP system is changed, which limits their adoption on operational forecasting chains (Zanchetta & Coulibaly, 2020). Boullange (reference) has developed a method based on observations that would not require the update of the statistical calibration of the NWP model; however, compared to ecPoint, it means using only one G\_WT and, while producing acceptable verification results, there is a benefit on using multiple G\_WTs that can be anticipated by the significant differences between the G\_WTs themselves (Hewson & Pillosu, 2021).