Highlights

# Optimizing warnings from hydrologic ensemble prediction systems for improved decision making

Jesu´s Casado-Rodr´ıguez, Peter Salamon, Stefania Grimaldi, Corentin Car- ton De Wiart, Christel Prudhomme, Calum Baugh, Ervin Zsoter, Ilias Pech- livanidis, Nina Bosshard, Michaela Mikulickova

* Probabilistic NWP models outperform deterministic counterparts in terms of flood warning skill.
* Using persistence (consecutive forecasts) as a warning criterion is not effective in probabilistic systems.
* Building grand ensembles adds value to the skill of the best performing NWP but this added value decreases with lead time.

Optimizing warnings from hydrologic ensemble prediction systems for improved decision making

Jesu´s Casado-Rodr´ıgueza, Peter Salamona, Stefania Grimaldia, Corentin Carton De Wiartb, Christel Prudhommeb, Calum Baughb, Ervin Zsoterb,

Ilias Pechlivanidisc, Nina Bosshardc, Michaela Mikulickovad

*aEuropean Commission, Joint Research Centre, Directorate E - Space, Security and Migration, Via Enrico Fermi 2749, Ispra, 21027, (VA), Italy*

*bEuropean Centre for Medium Range Weather Forecasts, Shinfield Park, Reading, WC2R 2LS, UK*

*cSwedish Meteorological and Hydrological Institute, Folkborgsv¨agen 17, Norrk¨oping, Sweden*

*dSlovak Hydrometeorological Institute, Jes´eniova 17, Bratislava, Slovakia*

# Abstract

Early Warning Systems are critical in mitigating the impacts of floods, the most frequent and economically damaging natural hazard worldwide. These systems rely on hydrological simulations driven by meteorological forecasts, both of which are inherently uncertain. Ensemble Prediction Systems (EPS) were developed decades ago to account for these uncertainties by generating multiple future scenarios that portray the forecast uncertainty. The advan- tages of probabilistic modelling have been proven over the last decades both in the fields of meteorology and hydrology. In order to issue a flood warning, the probabilities provided by EPS need to be converted into dichotomous values of ”warning” or ”not warning”. This conversion is often done by ap- plication of thresholds, whose definition is pivotal for the final skill of the system. While meteorological and hydrological models are under continuous development, the methodologies for converting discharge forecasts into flood warnings lack continuous evaluation. In this paper, we use discharge simulations from the European Flood Awareness System (EFAS) to assess the skill of the four Numerical Weather Prediction (NWP) models of different nature —probabilistic and deterministic—within the system, and search for methods to combine all the NWP into a grand ensemble that enhances overall skill. Additionally, we evaluate the effects of the current notification criteria —probability threshold

*Preprint submitted to Journal of Hydrology February 1, 2024*

and persistence— and determine their optimal values. Our results demon- strate that probabilistic NWP models outperform deterministic ones in terms of flood warning skill, which demonstrate skill only at short lead times (up to 3 days). Additionally, our findings sug- gest that the persistence criterion is beneficial for removing inconsistencies between consecutive forecasts in deterministic NWP, but should be omitted from probabilistic systems. By applying optimal notification criteria to a single probabilistic NWP (ECMWF-ENS), we observe an improvement of up to 6.4% in skill compared with the current EFAS procedure, which employs four NWP models. When compared with the optimized ECMWF-ENS warnings, grand ensembles yield significant improvements in skill (up to 4.2%) only within the first two lead days. Two methods to build the grand ensem- ble stand out —equiprobability among all model runs and skill weighted, and we discuss advantages and drawbacks of these two approaches. Overall, our study underscores the importance of selecting appropriate flood notification criteria and limits the value added to the grand ensemble by deterministic NWPs to the shortest lead time range.

*Keywords:* EFAS, early warning system, flood warnings, probabilistic forecasting

1 **1. Introduction**

2 Floods are the natural disaster that causes the biggest economical losses in

3 Europe [WMO (2021).](#_bookmark59) In the European Union (EU), flood damages amounted

4 to a total cost of approximately EUR 280 billion in the period 1980-2022,

5 representing 43% of all meteorological hazards [EEA](#_bookmark42) [(2023).](#_bookmark42) The historical

6 data show an increasing tendency in flood damages over Europe [EEA (2023);](#_bookmark42)

7 [Paprotny et al.](#_bookmark56) [(2018](#_bookmark56)), and climate projections indicate an exacerbation of

8 this trend with increasing frequency and severity of extreme events [Alfieri](#_bookmark32)

9 [et al.](#_bookmark32) [(2017);](#_bookmark32) [Martina Angela Caretta et al.](#_bookmark52) (2023). In this scenario, Flood

10 Early Warning Systems (FEWS) emerge as one of the multiple adaptation

11 measures to be taken. The major flood event in the Elbe River in 2002

12 triggered the development of the European Flood Awareness system (EFAS),

13 a continental scale FEWS that has been operational since 2012 as part of the

14 Copernicus Emergency Management Service and whose monetary value has

15 been proven [Pappenberger et al.](#_bookmark55) [(2015)](#_bookmark55). The magnitude of the flood that

16 hit Germany and Belgium in July 2021, the most costly in the last 40 years,

17 [EEA](#_bookmark42) [(2023)](#_bookmark42) was a reminder of the importance of FEWS and their need for

18 continuous improvement. On a global scale, the United Nations launched the

19 initiative Early Warnings for All to ensure that the entire world

20 is warned about weather, water and climate hazards by 2027 [Egerton et al.](#_bookmark43)

21 [(2023).](#_bookmark43)

22 FEWS rely on meteorological forecasts generated by Numerical Weather

23 Prediction (NWP) models. Over the past thirty years, Ensemble Predic-

24 tion Systems (EPS) have been developed within the realm of NWP models

25 to address uncertainties in the initial conditions. EPS provide equiprob-

26 able forecasts at any given time and location, enabling the quantification

27 and dissemination of probabilistic forecasts [Cloke and Pappenberger (2009).](#_bookmark39)

28 In comparison to deterministic models, EPS have demonstrated enhanced

29 consistency across consecutive forecasts [Buizza (2008)](#_bookmark37) and improved precip-

30 itation forecasts at longer lead times [Molteni et al. (1996).](#_bookmark53)

31 Approximately two decades ago, the success of EPS in meteorology led

32 to the adoption of this paradigm in flood forecasting [Diomede et al. (2008);](#_bookmark41)

33 [Bartholmes et al.](#_bookmark33) [(2009);](#_bookmark33) [Thielen et al.](#_bookmark58) [(2009);](#_bookmark58) [Yang et al.](#_bookmark60) [(2020).](#_bookmark60) Hydro-

34 logical Ensemble Predictions Systems (HEPS) propagate the uncertainties in

35 the meteorological forecasts by forcing hydrological models with the members

36 of a meteorological ensemble. Although this approach is the most common,

37 it only considers the uncertainty in the initial meteorological condition, but

38 does not include other sources of uncertainty, such as model structure

39 or model parameterization [Cloke and Pappenberger (2009).](#_bookmark39)

40 The decision to issue weather-related warnings, such as floods, is delicate

41 as it might result in significant costs, either due to unpreparedness if an event

42 is missed or the mobilization of resources in false alarms. Weather forecast-

43 ing systems often issue a higher number of false alarms than missed events

44 [Bouttier and Marchal](#_bookmark35) [(2023),](#_bookmark35) despite the potential risk of eroding trust in

45 the system, known as the ’cry wolf’ effect [LeClerc and Joslyn](#_bookmark50) [(2015).](#_bookmark50) This

46 latter research showed that an increase in false alarms diminishes the credi-

47 bility of the system, leading end-users to disregard precautionary measures.

48 However, they also demonstrated that incorporating uncertainty estimates in

49 warning messages can assist users in making informed decisions. Therefore,

50 HEPS have the potential to enhance both the skill and trustworthiness of

51 flood warning system, if particular attention is given to the balance between

52 false alarms and missed events.

53 EFAS is the hydrological forecast component of the Copernicus Emer-

54 gency Management Services (CEMS) of the European Commission [Thielen](#_bookmark58)

55 [et al.](#_bookmark58) [(2009).](#_bookmark58) EFAS is an operational system that monitors and forecasts

56 floods in an extensive European domain, encompassing not only European

57 Union member states. Its primary objective is to enable proactive measures

58 in anticipation of significant flood events. EFAS functions as a complemen-

59 tary system to national or regional counterparts, with a specific focus on

60 trans-boundary rivers and medium-range forecasts.

61 With the release of version 4.0, EFAS has been generating forecasts

62 twice a day, with a 6-hour resolution and lead times extending up to 10

63 days. It couples four weather forecasts of diverse nature —deterministic

64 and probabilistic— with the distributed, processed-based hydrological model

65 LISFLOOD-OS [Burek et al.](#_bookmark38) [(2013).](#_bookmark38) The simulation produces a set of wa-

66 ter fluxes and state variables, from which river discharge is utilized to issue

67 flood warnings. EFAS employs a threshold exceedance approach for issuing

68 warnings [Bartholmes et al.](#_bookmark33) [(2009),](#_bookmark33) where continuous discharge time series

69 are converted into binary series of exceedance or not-exceedance over a dis-

70 charge threshold (the 5-year return period). In their work, Bartholmes et

71 al. [Bartholmes et al.](#_bookmark33) [(2009)](#_bookmark33) developed a notification criteria based on two

72 types of persistence: persistence in a particular forecast is determined by the

73 number of NWP members predicting the exceedance, while persistence over

74 several forecast establishes that a flood signal is reliable when consecutive

75 forecast initializations predict the event. In this study, we will use the term ”probabil76 ity of exceedance” to refer to the first type of persistence, as the number of

77 NWP members exceeding the discharge threshold can be translated into a

78 probability, assuming equiprobability among them. We will keep the term

79 ”persistence” to denote only the consistency among consecutive forecast initializations.

80 The criteria applied in EFAS to issue flood warnings have changed over the past years driven by experience of the forecasters and feedback from users. Currently, a 30\% probability threshold and a persistence of 3 consecutive forecasts are used. These cri-

83 teria are independently applied to the two deterministic and the two proba-

84 bilistic NWP models, issuing a notification only if at least one of the models

85 of each type predicts the flood. Additionally, EFAS notifications are exclu-

86 sively dispatched to locations with a minimum catchment area of 2000 km²

87 and for lead times longer than 48 hours. The first limit is based on the infe-

88 rior hydrological performance in smaller catchments, while the second limit

89 is a consequence of EFAS addressing early awareness and serving as a supplementary system to national or

90 regional services. Previous studies have examined the skill of EFAS notifica-

91 tions [Bartholmes et al. (2009)](#_bookmark33) and the response of end users [Demeritt et al.](#_bookmark40)

92 [(2013).](#_bookmark40) However, the continuous advances in NWP performance [Bauer et al.](#_bookmark34)

93 [(2015),](#_bookmark34) as well as in the resolution and representation of hydrological pro-

94 cesses in EFAS [Joint Research Centre - European Commission (2020,](#_bookmark46) [2023),](#_bookmark47)

95 requires a new evaluation of the notification criteria to harness the potential

96 of the system.

97 The objective of this analysis is to assess the skill of a continental flood

98 forecasting system like EFAS in its current state, and explore new notifi-

99 cation criteria that can maximize its warning skill. In particular, the first

100 goal is to determine the value of combining several NWP models of different

101 nature in a single system, specifically seeking the appropriate combination

102 method to build a grand ensemble. The second goal is to optimize the two

103 notification criteria —probability threshold and persistence—, analysing the

104 interplay between them and reconsidering whether the persistence criterion is

105 meaningful in a purely probabilistic approach. The third and final goal is to

106 analyse the evolution of the notification skill with catchment area to discern

107 if the newer, higher resolution models are skillful in smaller catchments.

108 To address our research questions, we conducted an analysis of EFAS v.4

109 discharge simulations for the period 2020-2023. Section [2](#_bookmark0) enumerates the

110 datasets used in the analysis. Section [3](#_bookmark3) explains the two experiments that

111 analyse the skill of individual NWP models and different combinations of

112 those models, the criteria optimization process, and the skill metrics we tar-

113 geted. Section [4](#_bookmark14) follows a similar structure, presenting first the skill of in-

114 dividual NWP, followed by that of the combination methods, and ends with

115 an exploration of the relationship between notification skill and catchment

116 area. Section [5](#_bookmark31) discusses the importance of the spatio-temporal framework

117 and target metric in skill analysis like the one here presented, the trade-offs

118 between the notification criteria, the benefits of combining NWP models,

119 and the challenges in selecting a combination method. Finally, Section 6 states the conclusions.

120  **2. Data**

121 The data used in this analysis is taken from the operational setup of the

122 EFAS in its version 4 [Joint Research Centre - European Commission (2020).](#_bookmark46)

123 EFAS generates a hydrological forecast twice a day (00 and 12 UTC) forced

124 by the meteorological forecast of four different NWP models outlined in Table

125 [1.](#_bookmark1) EFAS integrates forecasts from probabilistic models (COS and ENS) and

126 deterministic models (DWD and HRES) to issue flood warnings.

127 The study period spans from the release of EFAS4 in October 2020 until

128 June 2023. For this period we use the EFAS forecasts as the discharge pre-

129 diction and the EFAS reanalysis as a proxy of discharge observations. Using

Table 1: Characteristics of the numerical weather prediction models used in EFAS4.

|  |  |  |
| --- | --- | --- |
| Model | Provider | Acronym Max. No. Spatial  lead time ensembles resolution |
| COSMO-LEPS |  | COS 5.5 days 20 ∼ 7 km |
| ICON-EU/ICON | DWD | DWD 7 days 1 ∼ 6.5-13 km |
| HRES | ECMWF | HRES 10 days 1 ∼ 9 km |
| ENS | ECMWF | ENS 10 days 51 ∼ 18 km |

130 reanalysis as ground truth addresses challenges driven by limited observational hydrological data (e.g.

131 gaps in the observed time series, difference in recorded period etc.), and removes

132 model representation and calibration errors from the analysis. The previous

133 data sets were converted from a continuous variable —discharge— into bi-

134 nary events of exceedance or not exceedance over a threshold. Following the

135 current EFAS setup, the flood threshold is the discharge associated with the

136 5 year return period, which was computed by fitting a Gumbel distribution

137 to the annual maxima from the discharge simulated in the EFAS4 long-term

138 run (1991-2023).

139 In contrast to the original EFAS skill assessment [Bartholmes et al. (2009),](#_bookmark33)

140 the current analysis focused solely on real gauging stations rather than all

141 river cells in the model, which would have cause an autocorrelation issue.

142 A selection was made from the approximately 4000 gauging stations in the

143 EFAS database based on two conditions: the contributing area must ex-

144 ceed 500 km², and the hydrological performance, in terms of the modified KGE

145 (Kling-Gupta efficiency) [Kling et al.](#_bookmark48) ([2012);](#_bookmark48) [Gupta et al.](#_bookmark44) [(2009);](#_bookmark44) [Knoben](#_bookmark49)

146 [et al.](#_bookmark49) [(2019),](#_bookmark49) must not fall below 0.50. The lower bound on catchment was

147 set due to the relatively coarse spatial resolution of NWP and EFAS temporal

148 resolution (6 h), which limits hydrological performance in small catchments.

149 The condition based on hydrological performance ensures the validity of us-

150 ing model simulations as proxies for observations is valid. Ultimately, 1979

151 gauging stations were included in the analysis, accounting for a total of 1683

152 events (exceedances over the 5 year threshold) during the study period. As

153 explained in section [3.4,](#_bookmark13) the optimization of the notification criteria was con-

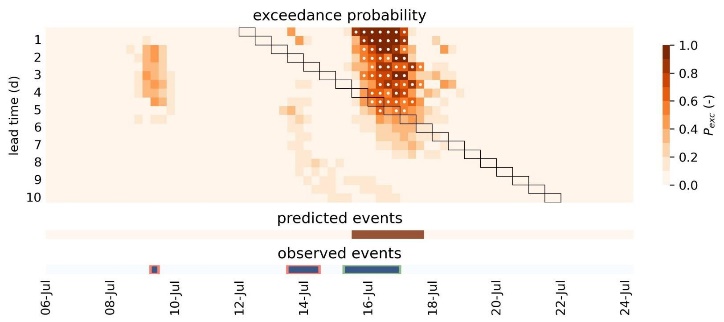
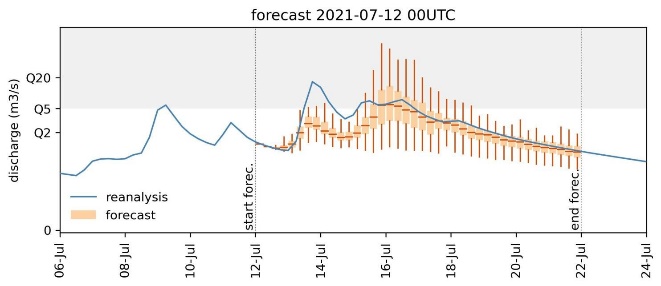
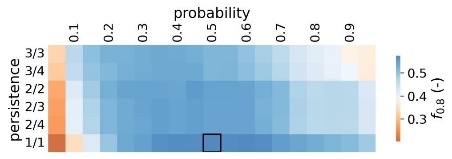
154 ducted on a subset of 1239 stations with a contributing area larger than 2000

155 km² (874 events), while the entire set was used to analyze the influence of

156 catchment area on warning skill.

|  |  |  |
| --- | --- | --- |
|  | predi  True | cted  False |
| True | 1 | 2 |
| observed  False | 0 | - |

Figure 1: Overview of the skill assessment framework. a) Hydrographs comparing observed and forecasted discharge (boxplot) with the discharge associated to the 5-year return period (Q5). b) Matrix of total exceedance probability (white dots identify pairs of date and lead time that meet the notification criteria, whereas black rectangles delineate the location of the forecast hydrograph from panel (a) in the probability matrix) and identification of predicted events for specific values of probability threshold and persistence. c) Confusion matrix for the selected station, constructed based on the previous notification criteria. d) Evaluation of the f-score across all combinations of probability threshold and persistence; the black square highlights the skill score corresponding to the notification criteria in panel (b).



a)

d)

b)

𝑃𝑟𝑜𝑏𝑎𝑏𝑖𝑙𝑖𝑡𝑦 ≥ 0.5

𝑃𝑒𝑟𝑠𝑖𝑠𝑡𝑒𝑛𝑐𝑒 = 1/1

c)

𝑓.0.8 = 0.56

157  **3. Methods**

158 The skill assessment comprises four steps, as depicted schematically in

159 Figure [1:](#_bookmark2) (a) transformation of both observed and forecasted discharge time

160 series into corresponding time series of the probability of exceeding the 5-

161 year return period; (b) construction of an exceedance probability matrix that

162 consolidates the overlapping forecasts, followed by identification of predicted

163 events through the application of specific notification criteria on that matrix;

164 (c) calculation of the contingency table —hits, misses and false alarms—, (d)

165 evaluation of the skill of the system across all possible combinations of the

166 notification criteria.

167 *3.1. Flood detection as exceedance over threshold*

168 The notification process starts by converting the continuous time series

169 of the target variable into time series of exceedance or non-exceedance over

170 a predefined threshold. Within EFAS, flood notifications are based on dis-

171 charge, and the critical threshold is the discharge associated with a 5-year

172 return period (*Q*5) generated from a 30-year historical simulation. While

173 alternative alert systems might rely on disparate variables, such as precip-

174 itation [Bouttier and Marchal](#_bookmark35) [(2023)](#_bookmark35) or river stage [Nevo et al.](#_bookmark54) [(2022),](#_bookmark54) and

175 employ different thresholds for defining events, the underlying methodology

176 remains similar.

177 The application of the discharge threshold yields distinct outcomes when

178 comparing a single time series—like observations or deterministic forecasts—with

179 an ensemble forecast. In the former case, the time series is converted into

180 a binary sequence, where 0 represents non-exceedance and 1 denotes ex-

181 ceedance. Conversely, ensemble forecasts lead to a probability time series,

182 with values fluctuating between 0 and 1. Since forecasts —both deterministic

183 and probabilistic— are generated more frequently (every 12 hours) than the

184 forecast horizon, they overlap. To address this, the ensemble forecasts are

185 restructured into an exceedance probability matrix that connects each time

186 step to increasing lead times, as illustrated in panel (b) of Figure [1.](#_bookmark2) Such

187 a configuration simplifies both the analysis and the subsequent computa-

188 tions. Within this matrix, an individual forecast is represented by a diagonal

189 (black rectangles). The uppermost rows represent shorter lead times, where

190 the probability of accurate prediction generally increases. For instance, in the

191 example in Figure [1,](#_bookmark2) the flood event occurring around July 16 was predicted

192 with a 50% probability five days in advance, increasing close to certainty

193 within two days of the event.

194  *3.1.1. Combination of NWPs*

195 One of the particularities of EFAS is the simultaneous use of four NWP,

196 both deterministic and probabilistic. The purpose of using multiple NWP is

197 to leverage the strengths of each model. Thus, a primary aim of this study

198 is to develop an effective method for integrating deterministic and proba-

199 bilistic forecasts. In practical terms, EFAS forecasts produce four distinct

200 probability matrices, one for each NWP, in contrast to the single matrix de-

201 picted in panel (b) of Figure [1.](#_bookmark2) The challenge lies in devising a methodology

202 to combine these four matrices into a total exceedance probability matrix,

203 thereby enhancing the system’s predictive accuracy compared to the use of

204 a single NWP. We explored three different approaches for calculating the to-

205 tal exceedance probability and compared them with EFAS current approach,

206 which does not incorporate a total probability framework.

207 In the current EFAS procedure, a notification is issued if at least one

208 deterministic and one probabilistic NWP predict the event. This approach

209 will be referred to as *1 deterministic + 1 probabilistic* (1D+1P) from here on.

210 As each model is evaluated independently, the current procedure does not

211 calculate a total probability matrix. The most straightforward method to

212 compute the total probability is the *model mean* (MM); it is a simple mean

213 over the four exceedance probability matrices, meaning all models are given

214 equal weight. In other words, the single member of a deterministic model is

215 considered significantly more important than any members of a probabilistic

216 model. An alternative approach would be to assign the same weight to each

217 member, regardless of whether it belongs to a deterministic or a probabilistic

218 model. This approach will be referred to as *member weighted* (MW). In this

219 approach, each NWP receives a weight relative to the number of members it

220 contains, meaning the probabilistic models have greater influence than the

221 deterministic ones. However, neither of these approaches take into account

222 the performance of the NWP. To address this limitation, the *Brier weighted*

223 approach (BW) assigns a weight to each model based on its probabilistic

224 skill. The Brier score [Brier (1950)](#_bookmark36) —a probabilistic error metric that ranges

225 from 0 to infinite, with 0 being the optimal value— is used as a measure of

226 probabilistic skill. The Brier score is calculated using the equation:

*T*

Σ1

BS = (*P T*

*t*=1

*obs,t*

* *Ppred,t*

)2 (1)

227 where *BS* is the Brier score of a single station, *T* is the number of time steps,

228 *Pobs,t* is the observed probability of exceedance, and *Ppred,t* is the predicted

229 probability of exceedance at a specific time step. The results of the BW

230 approach are highly sensitive to how Brier scores are converted into weights.

231 The conversion function must transform the values from a metric whose

232 optimal value is 0 to a weight with an optimal value of 1. Additionally, it

233 should be exponential to amplify the differences among NWP, considering the

234 extremely low BS values that are characteristic in rare events where both the

235 observed and predicted probabilities are 0 most of the time. Inverse distance

236 weighing (eq. [2)](#_bookmark6) fulfils both conditions. Values of the exponent *p* from 1 to

Table 2: Contingency table in a binary classification. *Eobs* is the sum of observed events and *Epred* the sum of predicted events.

Observed

|  |  |  |
| --- | --- | --- |
|  | True | False |
| Forecasted True | hit | false alarm *Epred* |
| False | miss | true negative |
|  | *Eobs* |  |

237

238

239

9 were tested, and 7 was found to be the optimal value.

*BS−p*

Σ

*w* = *BS*−*p* (2)

The weights in the BW method are specific to each NWP and lead time, whereas in the MM and MW methods they remain stable over lead time.

240 *3.2. Contingency table*

241 The skill assessment of a warning system involves a binary classification

242 task that compares two time series: the predicted and observed events. Typ-

243 ically, the contingency table (Table [2)](#_bookmark5) is used to summarize performance in

244 this type of classification problem. Given that floods are rare events, the

245 classification task is highly imbalanced, resulting in the number of true neg-

246 atives being orders of magnitude larger than any of the other terms in the

247 contingency table. Consequently, we must omit the true negatives from both

248 the computation of the contingency table and the selection of the target skill

249 metric not to overestimate skill (Section [3.3).](#_bookmark8)

250 The derivation of the binary time series of observed events is straight

251 forward by application of the discharge threshold (*Q*5) over the reanalysis

252 time series. Conversely, deriving the binary time series of predicted events

253 involves applying two notification criteria to the matrix of total probability

254 of exceedance (or the individual matrices of exceedance probability for each

255 NWP in the 1D+1P approach). The probability threshold sets the minimum

256 value of the exceedance probability at which a high risk of flooding is con-

257 sidered and a notification must be issued. In the current EFAS setup, the

258 threshold is 30% and applies only to the probabilistic forecasts. Addition-

259 ally, the persistence criterion was introduced to remove false alarms caused

260 by the erratic behaviour of some NWP [Bartholmes et al.](#_bookmark33) [(2009).](#_bookmark33) This cri-

261 terion aims to replicate the behaviour of a forecaster, who would wait to

262

263

264

265

266

267

268

269

270

271

272

273

274

275

276

277

278

279

280

send the warning until consecutive forecasts predict the event. In the cur- rent procedure, notifications are sent if 3 consecutive forecasts exceed the probability threshold (hereafter referred to as 3/3 persistence). Panel (b) in Figure [1](#_bookmark2) illustrates the application of these two notification criteria on the matrix of exceedance probability. For every time step (columns), the white dots identify the lead times that meet the criteria. Eventually, a flood event is predicted at a particular time step if any of the lead times meet the criteria. The terms of the contingency table result from comparing the predicted and observed events. We consider a hit any observed event in which at least one time step was correctly predicted. The number of misses and false alarms are calculated, respectively, as the difference between the observed (*Eobs*) or

predicted (*Epred*) events and the hits (eq. [3).](#_bookmark7)

misses = *Eobs* − hits (3)

false alarms = *Epred* − hits (4)

The process is repeated for all possible combinations of the two criteria and across all stations. We tested probability threshold values ranging from 5% to 95% and six persistence values: 1/1 (no persistence), 2/4, 2/3, 2/2, 3/4 and 3/3 (the current persistence criterion). The outputs of this process are matrices of hits, misses and false alarms for each station an combination of the probability threshold and persistence that will be used in the subsequent section to compute skill metrics.

281  *3.3. Skill metrics*

282 The fact that the classification task we face is highly imbalanced is fun-

283 damental in the selection of the skill metric from the myriad of options. For

284 instance, the odds ratio and the Hanssen-Kuipers score (*HK*) [Hanssen and](#_bookmark45)

285 [Kuipers (1965)](#_bookmark45) were discarded because both include the true negatives, which

286 is a term to be excluded in an imbalanced classification. Instead, we have

287 selected three skill metrics specific for imbalanced classification, i.e., that

288 exclude the true negatives.

289 Recall is the proportion of observed events that are correctly predicted

290 (eq. [5);](#_bookmark9) in some contexts it is also known as hit rate, probability of detection

291 or sensitivity. Bartholmes [Bartholmes et al.](#_bookmark33) ([2009)](#_bookmark33) proved that recall is equal

292 to HK for highly imbalanced classifications, such as floods. Precision, a.k.a.

293 positive predicted value or frequency of hits, is the proportion of the predicted

294 events that are correct (eq. [6).](#_bookmark10) There is a well known trade-off between recall

295 and precision. In the specific case of flood notifications, very strict criteria

296 would cause very few notifications with high certainty, which means that they

297 will be mostly correct (high precision), but a lot of events would be missed

298 (poor recall). On the other hand, very relaxed criteria would cause a large

299 amount of uncertain notifications, which means a small number of missed

300 events (high recall) but a large number of false alarms (poor precision). As

301 a compromise, the *fβ* score is a metric that balances recall and precision (eq.

302 [7).](#_bookmark11) In its most common version (*β* = 1) it is the harmonic mean of recall

303 and precision, granting equal importance to both terms. However, higher

304 importance can be given to precision (*β <* 1), therefore limiting the amount

305 of false alarms, or to recall (*β >* 1), reducing the amount of misses. In our

306 study, after suggestions from forecasters analyzing EFAS on a daily basis, we

307 selected *f*0*.*8 as the target skill metric. This metric balances precision and

308 recall giving a slightly higher value to the former. The idea behind is to limit

309 the amount of false alarms, which would jeopardize the trust of the recipients

310 of the warnings. Both recall, precision and the *fβ* score range from 0 to 1,

311 being 1 their optimal value.

312 As a final validation metric, we used bias, which represents the proportion

313 between the number of predicted and observed events and is calculated as

314 the quotient between recall and precision (eq. [8).](#_bookmark12) Its optimal value is 1,

315 which means that the number of notifications issued is equal to the number

316 of observed events, and it ranges from 0 to infinity. All four metrics presented

317 here can be plotted in the Roebber diagram [Roebber](#_bookmark57) [(200](#_bookmark57)9). We modified

318 the original Roebber diagrams and replaced the critical success index by our

319 target metric (*f*0*.*8).

320

recall =

precision =

hits

=

*Eobs*

hits

=

hits

hits + misses

hits

(5)

(6)

321

*Epred*

hits + false alarms

322

*f* = 1 + *β*2 precision · recall

*β*2 · precision + recall

*β*

(7)

bias = *Epred*

hits + false alarms recall

= =

(8)

323

*Eobs*

*3.4. Optimal notification criteria*

hits + misses

precision

324 The current EFAS notification criteria involve 5 variables: minimum

325 catchment area, minimum lead time, combination of NWP, probability thresh-

326 old and persistence. We analysed and optimized all these variables, initially

327 isolating each one to understand its individual effects.

328 We have conducted two main experiments to explore the benefits of com-

329 bining NWP. The first experiment analyzes each NWP individually with two

330 objectives: identifying the strengths and weaknesses of each NWP, and un-

331 derstanding how the notification criteria affect differently probabilistic and

332 deterministic NWPs. The second experiment compares the four combina-

333 tions of NWP explained in section [3.1.1.](#_bookmark4) The aim is to identify the most

334 appropriate combination method and compare it against the most skillful

335 NWP to assess whether the combination of NWP adds value to the system.

336 The setup of these two experiments is similar, apart from the difference in

337 the input data, and is explained in the following paragraphs.

338 For the majority of the analyses, we used only stations that met the

339 current minimum catchment area of 2000 km². With this subset of stations,

340 we explored the evolution of skill with lead time, probability and persistence.

341 The 20 lead time values —2 forecasts per day and an horizon of 10 days—

342 were grouped in two ways. First, we grouped them daily to analyze the

343 evolution of skill with lead time. Second, to explore the impacts of the

344 probability threshold and persistence, we defined 4 lead time ranges: a)

345 shorter than 2 days, a range at which EFAS formal notifications cannot be

346 issued, b) 2-5.5 days, when all NWP are available, c) 5.5-7 days, when 3

347 NWP are available as the maximum lead time of COS is exceeded, d) 7-10

348 days when only 2 NWP are available as the maximum lead time of DWD is

349 exceeded.

350 Following the exploration, we proceeded to optimize the probability thresh-

351 old and the persistence criteria. Specifically, we sought the combination of

352 probability and persistence that produces the maximum *f*0*.*8 at each of the

353 4 lead time ranges previously specified. The optimization was repeated for

354 every NWP individually and for every combination of the NWP. To ensure

355 a robust selection of criteria, we applied a 10-fold cross validation: 20% of

356 the stations were left aside as a test set, while the remaining 80% were sub-

357 divided in 10 folds. We computed the skill for every combination of 9 folds

358 and then averaged over folds. The optimal notification criteria were derived

359 from this average-over-fold skill matrix, as shown in panel (d) of Figure [1).](#_bookmark2)

360 Subsequently, the selected criteria were evaluated on the test set to prevent

361 overfitting.

362 Lastly, we explored the impact of catchment area on warning skill using

363 the entire set of stations. We calculated the notification skill for increasing

364 values of the minimum catchment area, ranging from 500 km² to 300,000

365 km², using the criteria previously optimized for the minimum catchment

366 area of 2000 km². Subsequently, we tested whether the skill of the system,

367 particularly that of small catchments, significantly improves with a more

368 complex notification criteria in which the probability threshold varies with

369 catchment area. We fixed the persistence criterion to that obtained in the

370 initial optimization and only tuned the probability threshold for increasing

371 values of the catchment area threshold. We compared the skill of the fixed

372 probability threshold with the skill of the area-specific probability threshold

373 to evaluate whether the new, more complex criteria pays off in terms of

374 performance.

375  **4. Results**

376 *4.1. Analysis of ”observed” events*

377 Figure [2](#_bookmark15) depicts the spatial distribution of the 1239 stations used in the

378 optimization of the notification criteria, i.e., those with a minimum contribut-

379 ing area of 2000 km² and a KGE not lower than 0.50. The colours represent

380 the amount of ”observed” flood events, while the histogram at the bottom

381 shows the frequency of events across the stations.

382 In total, the optimization of the notification criteria will be based on 1239

383 stations and 874 ”observed events”. A higher concentration of points with

384 events is observed in Central Europe, the British Isles, and Mediterranean

385 catchments. Significant flood events in the Rhine, Meuse, and Ebro are

386 visually evident on the map. Throughout the study period, 61% of the

387 stations did not exceed the 5 year return period, which means that hits and

388 misses can not exist, and the only term in the contingency table will be false

389 alarms. Consequently, the f-score for these stations can be either 0 or null.

390 Additionally, 29% of stations registered only one ”observed” event. These

391 nuances bear implications for the overall skill of the system, as we will explore

392 in subsequent discussions.

393 The imposition of a minimum contributing area of 2000 km² places a

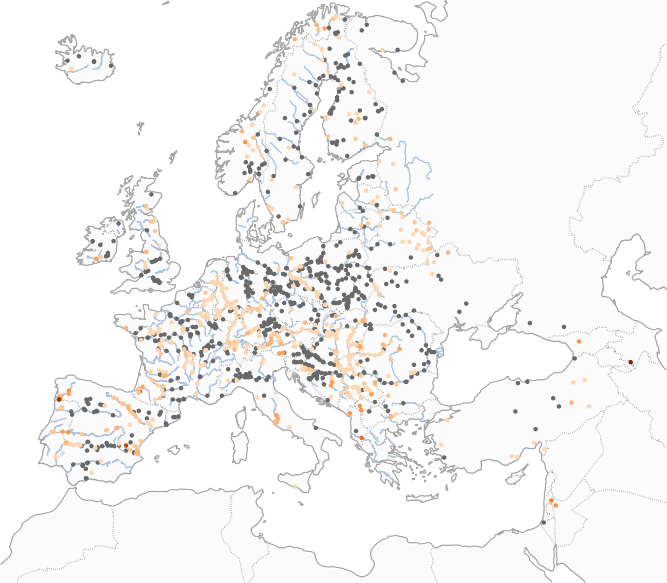
394 constraint on the number of stations available for optimization, consequently

395 affecting the count of observed events. In the final phase of our analysis, we

396 assessed the implications of this criterion by expanding our consideration to

397 stations with a minimum area of 500 km². Figure [3](#_bookmark16) represents the relation-

398 ship between an increasing catchment area threshold and the corresponding



1000

no. points

100

10

1

0

1 2 3 4 5 6 7 8 9 10 11 12 13

no. observed events

Figure 2: Geographical distribution of the stations used for optimizing the notification criteria, and number of ”observed” flood events during the study period.

2000

points

events

1500

1000

count (-)

500

current limit

0

103

104 105

area (km²)

Figure 3: Number of stations (orange) and observed events (blue) by catchment area threshold. The vertical, dotted line represents a catchment area of 2000 km², the current limit in the EFAS notifications.

399 number of stations and ”observed” events. Both counts decrease exponen-

400 tially with catchment area. By excluding the points with contributing area

401 smaller than 2000 km², we discarded 37% of the stations and 48% of the

402 events. The objective of this decision was to use the current notification

403 criteria as benchmark.

404 *4.2. Individual NWP*

405 Prior to studying how to combine the NWP models, we analysed the

406 skill of each of the models individually. This analysis serves a dual purpose.

407 First, to understand the relative skill of each model, with the intention of

408 incorporating this skill into the computation of the total probability. Second,

409 to compare the skill of individual NWP models against the benchmark—

410 the current notification criteria—, and to set a baseline for the combination

411 methods.

412  *4.2.1. Probabilistic skill assessment*

413 In the initial assessment of NWP skill, we calculated the Brier score for

414 each NWP at daily lead times. The Brier score, as an error metric, yields

415 values that can be challenging to interpret, given the dependence of error

416 magnitude on the specific variable—particularly in the context of rare events

417 like floods, where the magnitude is inherently low. Instead, we employ the

1.0

0.8

Brier skill score (-)

ENS COS HRES DWD

0.6

start notif.

end COS

end DWD

0.4

0.2

0.0

0 2 4 6 8

lead time (d)

Figure 4: Brier skill score of the four NWP used in EFAS at daily lead times. The reference (*BSS* = 0) is a model that never predicts an event. Blue lines represent probabilistic models, and orange lines deterministic ones.

418 Brier skill score (BSS) to gauge the relative skill of a model in comparison

419 to a reference. Models outperforming the reference receive a positive BSS,

420 while those underperforming register a negative BSS. For the reference, we

421 opted for a dummy model that never predicts an event (*Ppred* = 0) [Legg and](#_bookmark51)

422 [Mylne (2004).](#_bookmark51)

423 The plot above demonstrates that probabilistic models exhibit greater

424 skill than deterministic ones. Only within very short lead times (0-1 days)

425 do deterministic models approach the skill observed in probabilistic models.

426 This is particularly significant for EFAS notifications, as the formal noti-

427 fications are reserved for lead times greater than 2 days (indicated by the

428 leftmost dotted line). As lead time increases, there is a degradation in skill.

429 This decline impacts ENS to a lesser extent than the other models. Both de-

430 terministic models display poor skill at their forecast horizon. For instance,

431 at 10 days lead time, the skill of HRES is worse than a model that never

432 predicts a flood.

433 Overall, the most skillful model for flood warning forecasting is ENS,

434 while the least skillful is DWD. In section [4.3,](#_bookmark23) we will convert the Brier score

435 into weighing factors for the computation of total probability (see Figure [8).](#_bookmark24)

436 *4.2.2. Notification skill*

437 In this section we analysed how the NWP would have performed if notifi-

438 cations would be sent based uniquely in one of them. First, we illustrate the

439 general behaviour of notification skill according to the three dimensions here

1.0

0.8

0.6

f0.8 (-)

0.4

0.2

COS

start notif.

end COS

end DWD

DWD

ENS

HRES

current

P 0.10

P 0.20

P 0.30

P 0.40

P 0.50

P 0.60

P 0.70

P 0.80

P 0.90

0.0

0 2 4 6 8

0 2 4 6 8

0 2 4 6 8

0 2 4 6 8

lead time (d)

lead time (d)

lead time (d)

lead time (d)

Figure 5: Evolution of the notification skill with daily lead times and probability threshold for each NWP model in an scenario with no persistence. Each plot represents a different NWP. As a benchmark, the black, solid line represents the skill of the current operational criteria (1 deterministic + 1 probabilistic, 30% probability and a persistence of 3 forecasts).

440 involved —lead time, probability threshold and persistence. Second, we

441 optimize the probability and persistence criteria for a fixed lead time range.

442 Figure [5](#_bookmark19) exhibits the evolution of the notification skill with lead time and

443 probability threshold for a scenario with no persistence. The black, solid line

444 is the notification skill of the current notification criteria, as a benchmark.

445 Each colour line represents a different probability threshold; orange lines

446 represent thresholds below 50% and blue lines over that value. For the two

447 deterministic models (DWD, HRES) all lines overlap, since the probability

448 threshold does not affect a deterministic forecast.

449 In line with Figure [4,](#_bookmark18) the probabilistic models show overall higher skill than the

450 deterministic models. Under certain conditions, they even outperform the

451 current notification criteria, indicating potential for enhancing the skill of

452 the current EFAS notification criteria. Deterministic models demonstrate

453 skill on par with the current criteria solely for the initial lead time. As ex-

454 pected, notification skill degrades with lead time. This trend is consistent

455 across all cases —deterministic, probabilistic and benchmark—, although the

456 rate of degradation depends on the probability threshold. Lower probability

457 thresholds sharply degrade at short lead times, while high probability thresh-

458 olds experience a more pronounced loss of skill at longer lead times (such as

459 ENS over 7 days).

460 There is a range of probability thresholds, approximately from 40% to

461 90%, exhibiting similarly good skill. Within this range, higher thresholds

462 perform slightly better at short lead times, and lower thresholds perform

463 slightly better at mid-long lead times. As we are seeking to identify a single

464 optimal probability threshold for the notification criteria, the selection of the

1.0

0.8

0.6

f0.8 (-)

0.4

COS

DWD

ENS

HRES

current

1/1

2/4

2/3

2/2

3/4

3/3

0.2

0.0

0.2

0.4 0.6 0.8

probability

0.2

0.4 0.6 0.8

probability

0.2

0.4 0.6 0.8

probability

0.2

0.4 0.6 0.8

probability

Figure 6: Evolution of the notification skill with probability threshold and persistence for each NWP model at a fixed lead time range (from 2 to 5.5 days). Each plot represents a different NWP model. As a benchmark, the black cross represents the current criteria (1 deterministic + 1 probabilistic, 30% probability and a persistence of 3 forecasts).

465 optimal criteria will prioritize a lead time range from 2 days (the minimum

466 formal notification lead time) to 5.5 days (the maximum lead time in COS).

467 This approach allows us to mitigate the substantial uncertainty associated

468 with longer lead times, and focus on the lead times at which the majority of

469 formal notifications are issued and actions are taken from first responders.

HERE WE NEED A NEW SUB-SUBSECTION TO MAKE IT EASIER TO READ

470 Up to this point, we have not yet considered persistence in our analysis.

471 In order to assess the impact of persistence on the notification skill,

472 Figure [6](#_bookmark20) illustrates the evolution of skill with the

473 probability threshold and persistence for a fixed lead time range (2-5.5 days).

474 The benchmark (current notification criteria) is now represented by a single

475 point corresponding to a 30% probability and a persistence of 3/3. As the

476 probability threshold does not affect deterministic forecasts, the lines for the

477 two deterministic models (DWD and HRES) are horizontal.

478 Three out of four NWP models are able to replicate or improve the skill

479 of the current notification criteria at certain combinations of persistence and

480 probability. Only DWD is less skillful than the benchmark. The skill of

481 the probabilistic models shows a trade-off between probability threshold and

482 persistence. Both criteria play the same role of removing false positives (or

483 uncertain events) by taking stricter values. Consequently, stricter persistence

484 necessitates lower probability thresholds to optimize notification skill, and vice versa.

485 Nevertheless, the highest skill are consistently obtained in the scenario of no

486 persistence (1/1). In this scenario, there is a wide range of probability thresholds

487 (from 40 to 90%) that perform similarly well.

488 Persistence does have a significant impact in the deterministic models.

489 Their skill improves dramatically as soon as some persistence is added in the

Table 3: Summary of the optimization of the notification criteria for individual NWP and four lead time ranges. The initial row serves as a benchmark, indicating the skill of the current notification criteria. In each lead time range, bold fonts highlight the maximum obtained skill for a specific metric.

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| lead time | model | probability | persistence | recall | precision | f0*.*8 | rank |
| *<*2 d | 1D+1P | 0.300 | 3/3 | 0.522 | **0.735** | 0.634 | 3 |
|  | COS | 0.875 | 1/1 | 0.673 | 0.718 | **0.700** | 1 |
|  | DWD | - | 2/2 | 0.585 | 0.583 | 0.583 | 5 |
|  | ENS | 0.800 | 1/1 | **0.681** | 0.711 | 0.699 | 2 |
|  | HRES | - | 2/2 | 0.580 | 0.671 | 0.632 | 4 |
| 2-5.5 d | 1D+1P | 0.300 | 3/3 | 0.413 | 0.618 | 0.518 | 3 |
|  | COS | 0.725 | 1/1 | 0.387 | 0.687 | 0.518 | 2 |
|  | DWD | - | 3/3 | **0.420** | 0.522 | 0.477 | 5 |
|  | ENS | 0.650 | 1/1 | 0.416 | **0.725** | **0.562** | 1 |
|  | HRES | - | 3/3 | 0.416 | 0.612 | 0.517 | 4 |
| 5.5-7 d | 1D+1P | 0.300 | 3/3 | 0.215 | **0.612** | 0.356 | 2 |
|  | DWD | - | 2/2 | **0.291** | 0.358 | 0.328 | 4 |
|  | ENS | 0.450 | 1/1 | 0.264 | 0.605 | **0.402** | 1 |
|  | HRES | - | 2/2 | 0.273 | 0.429 | 0.351 | 3 |
| 7-10 d | 1D+1P | 0.300 | 3/3 | 0.120 | **0.625** | 0.237 | 3 |
|  | ENS | 0.175 | 2/2 | **0.287** | 0.396 | **0.345** | 1 |
|  | HRES | - | 2/2 | 0.244 | 0.336 | 0.293 | 2 |

490 criteria. For the most stringent persistence (3/3), the skill of the determin-

491 istic HRES is as good as that of the probabilistic ENS, and even better than

492 COS. However, as mentioned earlier, this is a sub-optimal value of persistence

493 for the probabilistic models.

494 With the knowledge gained in the previous exploration, we were in a bet-

495 ter position to understand the outcome of optimizing the notification criteria

496 individually for each NWP, summarized in Table 1. The results are organised

497 by lead time ranges; for each range, the table presents the optimized criteria

498 —probability and persistence— and the skill metrics —recall, precision and

499 f0.8— for the available NWP models. As a benchmark, the first row in each

500 lead time range shows the skill of the current operational criteria.

501 The current EFAS setup (1D+1P) notifies 52% of the actual events (re-

502 call) and 74% of the notifications are correct (precision) at most. These

503 figures reflect a negative bias, i.e., notifications are issued with a high level

504 of certainty, leading to a limited number of false alarms, but many events

505 being missed. This behaviour aligns with the selection of a biased target

506 metric, such as the *f*0*.*8 score, for the optimization of the notification criteria.

507 The current criteria is not optimal as 1D+1P is consistently outperformed in

508 terms of *f*0*.*8 at any lead time by single NWP, suggesting that the combination

509 of all NWPs hinders the skill ot that single NWP.

510 At the shortest lead time range, COS emerges as the top-performing

511 NWP, closely followed by ENS. Both ENS and COS optimized a very high

512 probability threshold and removed persistence. On the other hand, the skill

513 of the two deterministic NWP is distinctively lower —particularly DWD—,

514 and both require a persistence of 2/2. The current operational criteria is

515 outperformed by any of the probabilistic models, and the deterministic HRES

516 almost reaches that benchmark.

517 The scenario remains similar when examining lead times from 2 to 5.5

518 days. The two probabilistic NWPs continue to demonstrate higher skill than

519 the deterministic ones, and outperform the current operational criteria. In

520 this case, ENS is the top-performing NWP. Neither of them require persistence

521 and the probability thresholds are lower than in the previous lead time range,

522 although still higher than the 30% value used currently. HRES, despite being

523 deterministic, almost matches the skill of the benchmark and COS. As well

524 as DWD, to maximize its skill requires a persistence of 3/3. Overall, there

525 is a loss in skill of approximately 20% compared with the first 2 days of

526 forecast.

527 At lead times from 5.5 to 7 days, ENS

528 continues to outperform the other models. The optimal criteria follow the

529 trend: no persistence and a decreasing probability threshold, though still

530 higher than the benchmark. Overall skill has degraded by 30% compared

531 to the previous lead time range (over 40% compared to the first 2 days of

532 forecast).

533 The skill at the longest lead time range is overall poor. Only two models

534 have such a long forecast horizon, with ENS demonstrating significantly bet-

535 ter skill. This long lead time range is the only one for which ENS requires

536 persistence (2/2), a factor that notably decreases the probability threshold,

537 reflecting the trade-off between these two criteria.

538 The Roebber diagram in Figure [7](#_bookmark22) offers an alternate perspective on the

539 same results. Robber diagrams condense four skill metrics into a single plot

540 [Roebber](#_bookmark57) [(2009),](#_bookmark57) with an ideal model positioned at the top right corner.

541 Precision and recall are indicated on the X and Y axes, respectively. The

542 slope of the dashed lines represents bias, calculated as the ratio of recall

1.0

7.6

3.5

bias

2.1

1.4

0.8

0.6

recall

0.4

0.2

0.0

0.6

0.4

0.8

0.6

0.4

0.2

0.2

0.1

0.0 0.2 0.4 0.6 0.8 1.0

precision

current

NWP

ENS COS HRES DWD

bias

COMB

BW MW MM 1D+1P

Figure 7: Roebber diagram of the optimized criteria at the lead time range from 2 to 5.5 days. Four skill metrics are represented: precision and recall in the X and Y axis, bias by the dashed lines, and the *f*0*.*8 score by the solid lines. Triangles illustrate individual NWP and circles the combination of NWP. For reference, the skill of the current notification criteria is shown as a black cross.

543 to precision. Points below the 1:1 line indicate negative bias (a smaller

544 number of notifications than observed events), while points above the line

545 indicate positive bias (an excessive number of notifications). The *f*0*.*8 score

546 is represented by solid, black lines. Figure [7](#_bookmark22) shows simultaneously the skill

547 of the current notification criteria, the optimization of individual NWP, and

548 the optimization of combination methods (refer to Section [4.3)](#_bookmark23) for the lead

549 time range 2-5.5 days.

550 All NWP models and the current criteria exhibit negative bias, indicating

551 that EFAS underpredicts the occurrence of flood events. In the case of the

552 optimizations carried out in this analysis, the selection of the *f*0*.*8 score as the

553 target metric in the optimization forces this bias, as it prioritizes precision

554 over recall. However, the current setup already shows a negative bias. It is

555 remarkable that the improved *f*0*.*8 of the probabilistic models is a result of

556 higher precision, while recall values remain similar across all NWPs and the

557 current criteria.

558  *4.3. Combination of models*

559 As explained in the methodology, three of the four combination methods

560 compute total probability, which requires assigning weights to each NWP

561 and lead time. Figure [8](#_bookmark24) illustrates the distribution of weights among NWP

562 models for these three combination methods, and how the weights vary with

1.0

0.8

cumulative weight (-)

0.6

model mean

member weighted

Brier weighted

DWD HRES COS ENS

0.4

0.2

start notif.

end COS

end DWD

0.0

0 2 4 6 8

0 2 4 6 8

0 2 4 6 8

lead time (d)

lead time (d)

lead time (d)

Figure 8: Weights applied to compute the total probability. Every plot represents a different methods, with colours representing NWPs (blue for probabilistic and orange for deterministic models).

563 lead time.

564 The model mean (MM) approach attributes equal weights to each model.

565 The changes in the weights with lead time occur when the maximum lead

566 time of a model is exceeded (5.5 days for COS and 7 days for DWD). This is

567 the approach in which the deterministic models are given equal importance

568 as the probabilistic models.

569 Instead, in the member weighted (MW) combination, every model run

570 is assigned an equal weight. Therefore, ENS, which has a significant larger

571 number of members, prevails over any other NWP. Even for the first 5 days,

572 when all the models are available, ENS receives 70% of the weight. Con-

573 versely, the sum of the two deterministic models (HRES and DWD) has a

574 maximum weight of 4% from days 5.5 to 7. In general, this approach heavily

575 relies on the skill of the largest ensemble, which is not necessarily the most

576 skillful, as is in the case of EFAS.

577 In the Brier weighted (BW) combination, we used the Brier scores from

578 section [4.2.1](#_bookmark17) to derive weighing factors that prioritize the more skillful mod-

579 els. The resulting weights vary with lead time, showing a transition from

580 short to long lead times. At very short lead times, the deterministic models

581 demonstrate skill, accounting for approximately a third of the total weight.

582 However, as the lead time increases, the probabilistic models, particularly

583 ENS, assume the majority of the weight. At lead times from 2 to 5.5 days,

584 when most of the notifications are sent, the superior skill of probabilistic

585 models is represented by a combined weight of approximately 80%.

1 deterministic + 1 probabilistic

1.0

start notif.

end COS

end DWD

0.8

0.6

f0.8 (-)

0.4

0.2

model mean

member weighted

brier weighted

ENS

P 0.10

P 0.20

P 0.30

P 0.40

P 0.50

P 0.60

P 0.70

P 0.80

P 0.90

0.0

0 2 4 6 8

0 2 4 6 8

0 2 4 6 8

0 2 4 6 8

lead time (d)

lead time (d)

lead time (d)

lead time (d)

Figure 9: Evolution of the notification skill with daily lead times and probability threshold for each of the combination methods in a scenario with no persistence. Each plot rep- resents a different combination of NWP. As a benchmark, the black, solid line represents the skill of ENS with optimized criteria.

586 *4.3.1. Notification skill*

587 Figure [9](#_bookmark25) and Figure [10](#_bookmark26) replicate the analysis of the notification skill previ-

588 ously done for the individual NWP models, but this time for the combination

589 methods. Figure [9](#_bookmark25) illustrates the evolution of the notification skill with lead

590 time and probability threshold for a fixed persistence value (no persistence).

591 On the other hand, Figure [10](#_bookmark26) presents the evolution of skill with probability

592 and persistence for a fixed lead time range (2-5.5 days). In both cases, the

593 benchmark is ENS, the NWP that proved highest skill. Together these fig-

594 ures an overview of the behaviour of skill across four dimensions: combination

595 method, lead time, probability threshold and persistence.

596 The results indicate that ENS is responsible for the majority of the skill

597 in the grand ensembles, since only marginal gains are possible with specific

598 combinations at certain lead times. Only two combination methods (MW

599 and BW) are able to replicate the skill of the benchmark across all lead 600 times, with minimal improvements at days 2, 4 and 6. Although MM at 601 lead time 0 achieves the highest marginal gain, this analysis proves that this 602 combination method is not suitable, as the skill of this grand ensemble is 603 poorer than the benchmark from day 2 onwards. The current procedure 604 (1D+1P) is also sub-optimal, even though the degradation with lead time is 605 not as severe as with MM.

606 Regardless of the NWP or combination of NWPs, the notification skill

607 degrades notably with lead time. Even if at lead time 0 the skill is high (up 608 to 0.74), the minimum lead time criterion limits the EFAS notification skill 609 to values slightly lower than 0.6. The degradation continues and, from day 7 610 onward, the skill drops below 0.4, which may be an indicator for creating a

1 deterministic + 1 probabilistic

1.0

0.8

0.6

f0.8 (-)

0.4

model mean

member weighted

brier weighted

ENS

1/1

2/4

2/3

2/2

3/4

3/3

0.2

0.0

0.2 0.4 0.6

probability

0.8

0.2 0.4 0.6

probability

0.8

0.2 0.4 0.6

probability

0.8

0.2 0.4 0.6

probability

0.8

Figure 10: Evolution of the notification skill with probability threshold and persistence for each combination of NWP at a fixed lead time range (from 2 to 5.5 days). Each plot represents a different NWP model. Each plot represents a different combination of NWP. As a benchmark, the black cross represents the skill of ENS with optimized criteria (65% probability and no persistence).

611 new limitation on the maximum lead time at which notifications are issued. 612 Figure [10](#_bookmark26) reinforces the idea that the persistence criterion limits the 613 maximum attainable skill in ensemble systems. Across all four combina- 614 tions, the highest skill is reached when no persistence is applied (1/1). In 615 this scenario, all combinations exhibit similar maximum skill. Only member 616 weighted (MW) and Brier weighted (BW) marginally surpass the baseline 617 (ENS). These two combinations show optimal skill within a range of proba- 618 bility thresholds spanning 40% to 70%. In both instances, the maximum skill 619 is achieved at a probability threshold lower than that optimized for ENS.

620 The absence of sensitivity to probability in the current criteria (indicated

621 by the dark blue line in the left-hand side pane) is noteworthy. Even at 622 the lowest probability threshold, the *f*0*.*8 score remains at 0.5, unlike other 623 approaches where there is a significant decline in performance when the prob- 624 ability threshold takes very low values.

625 Table [4](#_bookmark27) summarizes the results of the optimization of the notification 626 criteria for the combination methods in an identical way as Table [3](#_bookmark21) did for 627 the NWPs. The results are organised by lead time ranges; for each range, the 628 table presents the optimized criteria (probability and persistence) and the 629 skill metrics (recall, precision and *f*0*.*8) for the four methods. As a baseline, 630 the first row in each lead time range shows the skill of ENS. The objective 631 of the optimization is again two-fold: to compare the four combinations of 632 NWP models, and to assess the added value of building a grand ensemble 633 instead of simply using the baseline.

634 In the shortest lead time range, two combinations clearly outperform

635 the baseline (MM, BW). Between the two, MM exhibits the highest skill

Table 4: Summary of the optimization of the notification criteria for the combination methods and four lead time ranges. The initial row serves as a baseline, indicating the skill of ENS, the top-performing NWP. In each lead time range, bold fonts highlight the peak value for a specific metric.

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| lead time | method | probability | persistence | recall | precision | *f*0*.*8 | rank |
| *<*2 d | ENS | 0.800 | 1/1 | **0.681** | 0.711 | 0.699 | 4 |
|  | 1D+1P | 0.925 | 1/1 | 0.662 | 0.717 | 0.694 | 5 |
|  | MM | 0.775 | 1/1 | 0.628 | **0.838** | **0.741** | 1 |
|  | MW | 0.750 | 1/1 | 0.673 | 0.718 | 0.700 | 3 |
|  | BW | 0.950 | 1/1 | 0.620 | 0.829 | 0.733 | 2 |
| 2-5.5 d | ENS | 0.650 | 1/1 | 0.416 | **0.725** | 0.562 | 3 |
|  | 1D+1P | 0.600 | 1/1 | **0.455** | 0.643 | 0.554 | 5 |
|  | MM | 0.700 | 1/1 | 0.424 | 0.700 | 0.559 | 4 |
|  | MW | 0.500 | 1/1 | **0.455** | 0.692 | 0.574 | 2 |
|  | BW | 0.525 | 1/1 | 0.451 | 0.700 | **0.576** | 1 |
| 5.5-7 d | ENS | 0.450 | 1/1 | 0.264 | **0.605** | **0.402** | 1 |
|  | 1D+1P | 0.425 | 1/1 | 0.262 | 0.599 | 0.399 | 4 |
|  | MM | 0.500 | 1/1 | **0.276** | 0.412 | 0.345 | 5 |
|  | MW | 0.425 | 1/1 | 0.271 | 0.579 | 0.401 | 2 |
|  | BW | 0.450 | 1/1 | 0.268 | 0.586 | 0.400 | 3 |
| 7-10 d | ENS | 0.175 | 2/2 | **0.287** | 0.396 | **0.345** | 1 |
|  | 1D+1P | 0.250 | 1/1 | 0.256 | 0.414 | 0.334 | 4 |
|  | MM | 0.225 | 2/2 | 0.252 | 0.334 | 0.296 | 5 |
|  | MW | 0.175 | 2/2 | 0.278 | 0.402 | 0.343 | 3 |
|  | BW | 0.300 | 1/1 | 0.255 | **0.445** | **0.345** | 1 |

636 (*f*0*.*8 = 0*.*741). This proficiency can be attributed to the importance that 637 MM assigns to deterministic models, which demonstrated notable skill at 638 very short lead times (see Figure [8).](#_bookmark24) As of the optimized criteria, none of 639 the four methods required persistence, and probability thresholds fell within 640 the top quartile.

641 From day 2 to 5.5, BW and MW stand out as the top-performing com- 642 binations, with a comparable skill that surpass the baseline by 1%. Both 643 approaches operate optimally without the need for persistence, with the prob- 644 ability threshold hovering around 50% (refer to Figure [10](#_bookmark26) for a sense of the 645 uncertainty in defining the probability threshold). Compared with the pre- 646 ceding lead time range, there is an evident decline in skill by approximately 647 20%.

648 In the third lead time range, from 5.5 to 7 days, none of the combinations

649 outperform the baseline, despite three of them (1D+1P, MW and BW) at- 650 taining very similar skill. None of these three methods necessitated the use 651 of persistence, and the associated probability thresholds were slightly lower 652 than those in the preceding lead time range (around 40%). When compared 653 with the prior lead time range, there is a decrease in skill of around 30% 654 (45% when compared to the the first range).

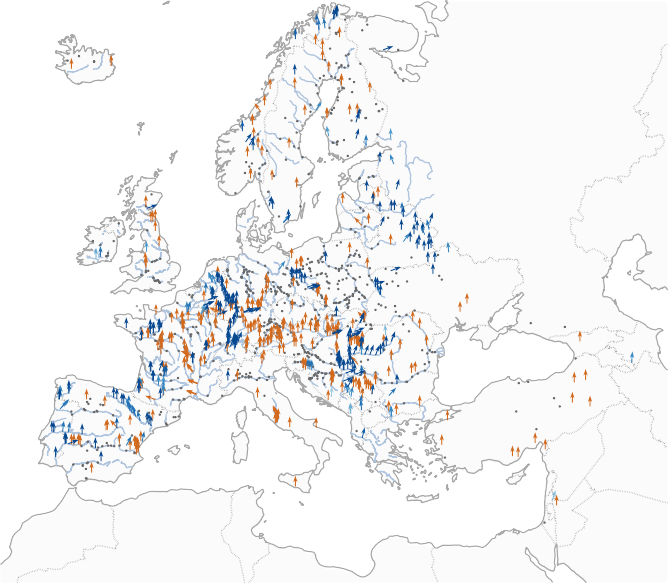
655 In the longest lead time range, BW emerges as the best-performing com- 656 bination, with a skill on par with the baseline. MW, while slightly less 657 proficient, shows similar skill; however, their optimized criteria differ. MW 658 requires a persistence of 2/2 and a relatively low probability threshold (0.175) 659 — precisely the same criteria optimized for the baseline. This results are not 660 surprising, as at this lead time range MW and ENS are practically the same, 661 since the weight of HRES in the grand ensemble is negligible. In contrast, 662 BW does not require persistence, and its probability threshold (30%) aligns 663 with the decreasing trend observed in previous lead time ranges.

A NEW SUB-SUBSECTION IS NEEDED HERE TO MAKE IT EARIER TO READERS TO FOLLOW

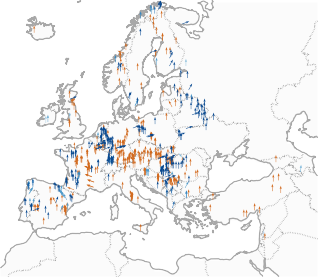
664 The maps in Figure [11](#_bookmark28) illustrate the skill metrics associated with the

665 optimal notification criteria, i.e., the BW combination with no persistence 666 and a probability threshold of 52.5%. The metrics represent only the lead 667 time range from 2 to 5.5 days, which corresponds to the period when the 668 majority of the notifications are issued.

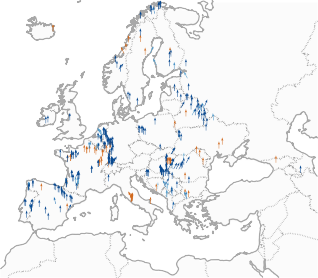
669 A pronounced contrast is evident in the maps, attributable to the limited 670 number of events in most stations, resulting in skill metrics primarily yielding 671 values of either 1 or 0. In terms of f-score, there is a clustering of under- 672 performing points in central Europe (notably the Alps) and central France. 673 The breakdown into recall and precision reveals that the predominant cause

*f*0.8

recall



precision



0 0.2 0.4 0.6 0.8 1.0

skill (-)

-0.5 0

-1



skill (-)

0.5

1

Figure 11: Skill maps for the optimized notification criteria (BW, 52.5% probability and no persistence) at the lead time range from 2 to 5.5 days. The main plot shows the skill in terms of *f*0*.*8 (the target metric); in this plot, gray dots are stations for which the f- score cannot be computed due to all instances of hits, misses and false alarms being null. The smaller plots show precision and recall, the components of the f-score; for the sake of visibility, it only shows the stations for which the specific metric can be computed. In all the plots the direction of the arrows shows the difference in that metric between the optimized criteria and the current criteria; positive values imply gains and negative values losses in skill.

1.0

0.8

0.6

skill (-)

f0.8

recall

precision

current 1D+1P MM MW BW

0.4

0.2

min. area

0.0

103

104 105

area (km²)

103

104 105

area (km²)

103

104 105

area (km²)

Figure 12: Evolution of skill with catchment area limit. Each plot represents a skill metric, with *f*0*.*8 as the optimization target, and recall and precision as its components. In each of the plots, the black line represents the skill of the current notification criteria, while the colour lines depict the skill of the optimized criteria for the 4 combination methods. The vertical, dotted line indicates the catchment area limit currently applied in EFAS.

674 of the low f-score lies in recall, indicating missed events. Conversely, the 675 precision of notifications remains consistently high, with exceptions observed 676 in the Seine and Arno rivers. It is noteworthy the skill in all metrics for 677 the Ebro, Rhine and Meuse—three rivers that experienced significant flood 678 events during the study period.

679 By applying this set of criteria, EFAS would be able to issue notifications 680 with at least 2 days lead time for 45% of the events, and 70% of the notifica- 681 tions would be correct. For reference, these figures represent an improvement 682 in both metrics over the existing criteria, which achieved 41% recall and 62% 683 precision (see Figure [7).](#_bookmark22) The Roebber diagram indicates that the grand en- 684 sembles —namely MW and BW— still suffer from a negative bias, but it was 685 slightly reduced by improving recall compared with ENS.

686 *4.4. Identification of optimal catchment area to issue notifications*

687 The preceding sections presented results for a set of stations with a catch- 688 ment area of at least 2000 km², the current limit. To investigate the potential 689 relaxation of this criterion, Figure [12](#_bookmark29) illustrates the evolution of skill depend- 690 ing on the catchment area limit. The criteria used to create this figure are 691 those optimized for the fixed area limit of 2000 km² and a lead time range 692 between 2 and 5.5 days (refer to Table [4).](#_bookmark27) For benchmarking purposes, the 693 plots compare the skill of the four combination methods (coloured lines) with 694 the skill of the current notification criteria (black line).

695 The skill of the system improves with catchment area, as expected. For 696 very large catchments, all three metrics reach their maximum value of 1. 697 However, the skill curves do not increase continuously. Recall experiences a 698 loss in skill for areas ranging from 30,000 to 70,000 km², leading to a decrease 699 in *f*0*.*8 values. This loss is caused by missed events in the Guadiana, Seine, 700 Loire, Rhone and Danube catchments. As already mentioned in previous 701 sections, there is a notable gap between recall and precision, the components 702 of the f-score, in both the optimized and current notification criteria, with 703 precision consistently surpassing recall.

704 The vertical line at 2000 km² illustrates the gain in skill achieved through 705 optimization, which is sustained across all catchment area values. Reducing 706 the catchment area limit results in minimal loss in *f*0*.*8, suggesting that this 707 criterion could be reduced without compromising skill. For instance, the BW 708 approach at 1,000 km² exhibits slightly higher skill (*f*0*.*8 = 0*.*531) than the 709 current criteria at 2000 km² (*f*0*.*8 = 0*.*518). The relaxation of the catchment 710 area limit impacts the f-score components differently. While precision re- 711 mains unaffected, there is a loss in recall. To sum up, issuing flood alerts to 712 smaller catchments would marginally reduce the skill of the system, primarily 713 due to an increase in missed events, without leading to more false alarms. This sets an opportunity to EFAS to increase its usability in even smaller river systems across Europe.

714 In the previous analysis, a fixed probability threshold was applied for all 715 catchments, regardless of their drainage area. We also investigated whether 716 the skill of the system could be enhanced by adapting the probability thresh- 717 old to the catchment area. Our rationale is rooted in the understanding 718 that the level of uncertainty (model spread) differs between small and large 719 catchments, suggesting that area-specific probability thresholds may enhance the 720 overall system skill. The results of this experiment are depicted in Figure 721 [13.](#_bookmark30) Each plot represents a different combination method, comparing the f- 722 score (solid lines) and the probability threshold (dashed lines) between the 723 model optimized for a 2,000 km² area limit (blue) and models optimized for 724 increasing catchment area limits (orange). In order to compare the results 725 of the two optimization runs, we kept a constant persistence value of 1/1, 726 which was identified as optimal in Table [4.](#_bookmark27)

727 The skill measured as *f*0*.*8 is very similar for both optimizations, especially 728 for areas smaller than 2,000 km². For larger areas, the main improvement 729 is at the range between 30,000 and 100,000 km², for which the fixed criteria 730 loses some skill. The evolution of the probability threshold with catchment 731 area is erratic, likely due to the uncertainty in defining the optimal value 732 (refer to Figure [10).](#_bookmark26) Despite this, a general trend can be discerned, with

1 deterministic + 1 probabilistic

1.0

persistence = 1/1

0.8

0.6

skill

model mean

member weighted

brier weighted

1.0

persistence = 1/1

0.8

0.6

f0.8 (optimal)

f0.8 (area optimized) prob. (optimal)

prob. (area optimized)

probability

0.4 0.4

persistence = 1/1

persistence = 1/1

0.2 0.2

0.0

103

104 105

area (km²)

103

104 105

area (km²)

103

104 105

area (km²)

103

104 105

area (km²)

0.0

Figure 13: Comparison of the notification skill between a fixed (blue) and a area-specific (orange) probability threshold criterion for the lead time range from 2 to 5.5 days. Each plot depicts a different combination method, with solid lines representing the *f*0*.*8 and dashed lines representing the probability threshold. The black, dotted, vertical line in- dicates the catchment area limit (2,000 km²) used for optimizing the fixed probability threshold.

733 the threshold taking lower values for catchments smaller than 2,000 km², 734 peaking at approximately 100,000 km², and dropping to a minimum of 5% 735 for very large catchments.

736 Overall, the experiment indicates that varying the probability threshold 737 does not significantly improve the skill of the notification criteria. Therefore, 738 for the sake of simplicity, a fixed probability threshold is recommended.

739  **5. Discussion**

5.1. Limitations and moving forward

740 The definition of the spatial and temporal framework is often a limitation 741 in studies that analyze the skill of ensemble flood forecasting systems [Cloke](#_bookmark39) 742 [and Pappenberger](#_bookmark39) [(2009).](#_bookmark39) Due to the rarity of flood events, the statistical 743 robustness of such analysis relies on a limited and often highly correlated 744 sample of events. To address this limitation, long-term and continental or 745 global studies are advocated, such as the one presented here, which spans 746 over a period of two and a half years and extends over a European domain. 747 Nevertheless, our study would benefit from a larger sample of points and flood 748 events. As shown in Figure [11,](#_bookmark28) there are many stations for which the skill 749 metrics cannot be calculated due to null hits, misses or false alarms. A larger 750 sample of observed events would reduce the susceptibility of the analysis to 751 this issue. Two potential solutions emerge: extending the study period, which 752 is limited by the availability of the HEPS forecasts (or reforecasts), or using 753 an event threshold of higher recurrence. We decided to work with a sample 754 of stations, instead of all the river cells [Bartholmes et al.](#_bookmark33) [(2009),](#_bookmark33) to limit 755 the spatial correlation of events. Additionally, we tested removing stations

756 with highly correlated discharge time series to further reduce correlation, but 757 the results were similar and the sample of events much smaller; hence, this 758 approach was discarded.

759 In the optimization of the notification criteria, we set a catchment area 760 limit of 2,000 km², a threshold inherited from previous EFAS setups with 761 lower spatial and temporal resolution. The results of this study (Figure 762 [12)](#_bookmark29) demonstrate that the current resolution of both NWP and hydrological 763 models can produce skillful warnings at smaller catchments, which can signif- 764 icantly increase the number of stations and flood events analyzed, as shown 765 in Figure [3.](#_bookmark16) In line with this, the new EFAS version 5 [(Joint Research Centre](#_bookmark47) 766 [- European Commission](#_bookmark47) [(2023))](#_bookmark47) has increased spatial resolution from 5 km 767 to 1 arc-minute, which is expected to further improve the skill of the system 768 in smaller catchments. The skill of this new EFAS version will be analyzed 769 as soon as a long enough period or reforecasts are available.

770 The selection of the target metric is a critical aspect in any skill analy-

771 sis. [Bartholmes et al.](#_bookmark33) (2[009)](#_bookmark33) reviewed the desired characteristics of a skill 772 score, and selected odds ratio, HK score and bias as their target metrics. 773 In contrast, for our analysis, we selected the *fβ* score, a metric commonly 774 used in machine learning, but rarely applied in hydrology. The *fβ* score has 775 two interesting properties for analyzing a flood warning system. Firstly, it is 776 based on the contingency table but avoids using true negatives, which could 777 overestimate skill. Secondly, it allows for tailoring the notification criteria to 778 the specific use case, as the *β* parameter can be adjusted to penalize stronger 779 misses or false alarms. For instance, Bouttier and Marchal [Bouttier and](#_bookmark35) 780 [Marchal](#_bookmark35) [(2023)](#_bookmark35) used the *f*2 score to optimize the probability threshold for 781 high-intensity precipitation warnings in France. The *β* value of 2 penalizes 782 four times more missed events than false alarms, under the assumption that 783 the cost of being hit by an extreme event unprepared is larger than the cost 784 of mobilising resources in a false alarm. In our study, we took the opposite 785 approach and decided to minimize the number of false alarms by using a *β* 786 value of 0*.*8. This decision was based on the fact that EFAS complements 787 the national or regional systems by issuing medium-range flood warnings to 788 the responsible administrations, not to the general public. The aim is to 789 alert authorities only when there is substantial evidence that the event will 790 occur, in order to avoid excessive notifications that could undermine their 791 trust in the system. Additionally, the event threshold (5-year return period) 792 corresponds to a recurrence time shorter than the flood protection levels in 793 many European regions, so we assume that missing minor events will not be

794 as costly as Bouttier and Marchal argue. In a preliminary analysis, we tested 795 *β* values of 0.8, 1 and 1.25 and found that the changes in the f-score were 796 minimal (in the order of 3%) and in none of the cases persistence was needed. 797 However, the selection of *β* affects the optimal probability threshold and the 798 resulting bias, i.e., the distribution of errors between misses and false alarms. 799 Values lower than 1 lead to higher probability thresholds and negative biases, 800 whereas values larger than 1 yield positive bias, i.e., more notifications than 801 actual events.

5.2. Addressing study objectives

802 The first objective of this study was to evaluate the skill of the differ-

803 ent NWP within EFAS, with a specific focus on comparing deterministic and 804 probabilistic models. Figure [4](#_bookmark18) demonstrates that probabilistic NWPs outper- 805 form deterministic models in terms of Brier skill score across all lead times, 806 with the exception of short lead times, where deterministic NWPs approach 807 similar levels of skill. The outcome is similar when applying the notification 808 criteria and measuring warning skill in terms of *f*0*.*8 (Figure [5](#_bookmark19) and Figure [6).](#_bookmark20) 809 At their optimal notification criteria, the probabilistic NWPs, particularly 810 ENS, outperform deterministic NWPs. The skill of both types of NWP de- 811 grades with increasing lead time, casting doubt on their ability to issue flood 812 alerts more than 5-6 days in advance. The effects of the probability thresh- 813 old and persistence criteria differs depending on the nature of the NWP. 814 The probability threshold has a substantial effect on the skill of probabilistic 815 models, while it does not impact deterministic models. Both COS and ENS 816 perform better at probability thresholds over 50%, with a wide range of sim- 817 ilarly performing values. On the contrary, persistence proves to be beneficial 818 in improving the notification skill of deterministic NWPs, but it limits the 819 skill of probabilistic models. Persistence, a.k.a. poor man’s ensemble, is a 820 method of creating an ensemble out of a deterministic forecast [Cloke and](#_bookmark39) 821 [Pappenberger (2009).](#_bookmark39) For instance, a persistence of 3/3 is an ensemble of 3 822 members with a probability threshold of 100%. Our results show that while 823 persistence removes the inherent erratic behaviour of deterministic models 824 and notably improves their skill, it should not be applied to HEPS.

825 The second objective of this study was to determine the appropriate

826 method to combine several NWP models into a grand ensemble. Careful 827 consideration is necessary in selecting this method to avoid diminishing the 828 skill of the most proficient NWP within the grand ensemble. For instance, 829 the current EFAS notification criteria limit the skill of the system to that 830 of the highest-performing deterministic model (HRES), which is poorer than 831 any of the probabilistic NWP (Figure [6).](#_bookmark20) This limitation is caused by a com-

832 bination method (1D+1P) that excessively relies on deterministic forecasts, 833 as well as the use of persistence. The challenge lies in finding a combination 834 that enhances the skill of the best-performing NWP, which in the case of 835 EFAS is the ENS.

836 Our findings demonstrate that ENS is a very strong baseline, with only a 837 few combination methods yielding gains to its skill. The effectiveness of the 838 grand ensemble is limited to the first 5 days lead time, when the total number 839 of members is highest (73), particularly for lead times shorter than 2 days, 840 when the deterministic models provide value. As an example, a simplistic 841 approach like MM exhibits the highest skill within the shortest lead time 842 range, potentially because it is the approach that assigns higher value to 843 the deterministic models. Overall, the most promising approaches are the 844 member-based (MW) and skill-based (BW) methods, as they outperform the 845 baseline throughout the first 5 days lead time. Both methods have similar 846 optimal notification criteria (no persistence and a probability threshold close 847 to 50%), and their success is based on granting most of the weight to ENS. 848 However, the reason why they allocate this weight to ENS differs. In the 849 case of MW, it is coincidental that the most skillful model also possesses 850 the largest number of members. On the contrary, BW grants greater value 851 to ENS because it proved to be a superior model. If a more skillful model 852 with fewer members would be added to the grand ensemble, BW would be 853 able to extract that value, while MW would not. Therefore, we argue that 854 skill-based combination methods are preferable.

5.3. Towards operationalization

855 The decision on the most appropriate combination methods can not be

856 based solely on skill metrics, but also consider other factors such as the opera- 857 tional implementation effort, user comprehensibility, or the impact of adding 858 or removing a NWP. Skill-based approaches, such as BW, are theoretically 859 superior as the allocation of weights is based on quality rather than quantity. 860 They are also the most robust when it comes to adding or removing NWP 861 models. However, they are more challenging to implement and to explain to 862 end users. To derive the matrix of weights (Figure [8),](#_bookmark24) a skill-based approach 863 requires having an extensive period of reanalysis and reforecasts upfront, 864 posing a substantial computational burden on system implementation. Ad- 865 ditionally, practitioners using the forecasts would require additional training 866 to understand that the ensemble members are not equiprobable, which would 867 be more intuitive, but there are some prevailing members.

868 There remains a question on the use of deterministic models in a grand

869 ensemble dominated by a probabilistic NWP. When using a member-based

870 combination, such as MW, deterministic models could be removed from the 871 grand ensemble, as the weight allocated to them is so small that their in- 872 clusion barely affects the total probability. Skill-based approaches, instead, 873 allocate larger weights to these models in the very first lead times, when 874 they have proven value given their often higher resolution. The inclusion of 875 deterministic models is a topic for future analysis, as the spatial resolution of 876 deterministic and probabilistic NWP are nowadays similar, and centres such 877 as the ECMWF plan to discontinue their deterministic models in favour of 878 the ensembles.

879 **6. Conclusions**

880 This study describes a procedure to optimize warning thresholds in HEPS. 881 We investigated the ability of deterministic and probabilistic NWP mod- 882 els in providing correct flood warnings, and we explored the effects of the 883 notification criteria, namely probability threshold and persistence. The fi- 884 nal objective of the study is to device a method to combine several NWP 885 into a grand ensemble in order to outperform the most proficient, individual 886 NWP. Our study case is the European Flood Awareness System (EFAS v.4), 887 a medium-range continental flood warning system.

888 The individual analysis of the NWPs demonstrated that probabilistic

889 models, particularly ENS, provide better flood warnings than deterministic 890 models, even for short lead times, where the latter show some skill. We 891 discovered that the persistence criterion is useful to increase the skill of de- 892 terministic models, as it creates a pseudo-ensemble, but is counterproductive 893 in the case of ensembles. The criterion that maximizes ensemble skill is the 894 probability threshold, whose optimum shows a certain degree of uncertainty, 895 but always exceeds a value of 50%.

896 By default, the combination of several NWP into a grand ensemble does

897 not improve the overall skill of the flood warning system. As an example, the 898 current EFAS notification criteria, which uses four NWPs, is outperformed 899 by only using one NWP (ENS) with the adequate criteria. Two of the ap- 900 proaches tested in this work yielded gains in skill compared with the baseline 901 in medium and short lead times. One of these approaches (named member 902 weighted) assumes equiprobability among all the model runs, and the other 903 one (named Brier weighted) assigns weights to each NWP at every lead time 904 based on their probabilistic skill. In our study case, these two approaches 905 yielded similar results because the most proficient NWP is also the one with

906 the largest number of members. Despite that, we argue that the skill-based 907 approach is conceptually more appropriate, gives significance to the inclusion 908 of deterministic NWP in the grand ensemble, and would perform better if 909 the set of NWP would change. However, other study cases may consider 910 other factors such implementation costs or explainability in the selection of 911 the combination method.

912 Finally, we explored how catchment area affects the notification skill, 913 with the idea of relaxing the current notification criterion that establishes a 914 minimum catchment area. The results indicate that skill deteriorates with 915 decreasing catchment area, as other studies report, but the improvements 916 in meteorological and hydrological models over the last decade allow for a 917 reduction in the minimum catchment area. We have proven that with the 918 optimized criteria we can halve the area limit, therefore cover a larger area, 919 and achieve a similar performance as in the current system setup. We expect 920 that the skill at smaller catchments will be positively affected by the release 921 of the new EFAS version with a higher spatial resolution.

# 922 7. Data availability

923 The reanalysis and forecast discharge data is available at the [Climate](https://cds.climate.copernicus.eu/) 924 [Data Store of the Copernicus Climate Change Service.](https://cds.climate.copernicus.eu/) All the scripts, pre- 925 processed data and results of the study can be found in this [GitHub reposi-](https://github.com/casadoj/EFAS_skill) 926 [tory.](https://github.com/casadoj/EFAS_skill)

927 **References**

928 Alfieri, L., Bisselink, B., Dottori, F., Naumann, G., de Roo, A., Salamon, 929 P., Wyser, K., Feyen, L., 2017. Global projections of river flood risk in a 930 warmer world. Earth’s Future 5, 171–182. doi:[10.1002/2016EF000485](http://dx.doi.org/10.1002/2016EF000485).

931 Bartholmes, J.C., Thielen, J., Ramos, M.H., Gentilini, S., 2009. The eu- 932 ropean flood alert system efas – part 2: Statistical skill assessment of 933 probabilistic and deterministic operational forecasts. Hydrol. Earth Syst. 934 Sci 13, 141–153. URL: [www.hydrol-earth-syst-sci.net/13/141/2009/](http://www.hydrol-earth-syst-sci.net/13/141/2009/),

935 doi:[10.5194/hess-13-141-2009](http://dx.doi.org/10.5194/hess-13-141-2009).

936 Bauer, P., Thorpe, A., Brunet, G., 2015. The quiet revolution of numerical

937 weather prediction. Nature 525, 47–55. doi:[10.1038/nature14956](http://dx.doi.org/10.1038/nature14956).

938 Bouttier, F., Marchal, H., 2023. Probabilistic short-range forecasts of high 939 precipitation events : optimal decision thresholds and predictability limits. 940 URL: <https://doi.org/10.5194/egusphere-2023-3111>, doi:[10.5194/](http://dx.doi.org/10.5194/egusphere-2023-3111) 941 [egusphere-2023-3111](http://dx.doi.org/10.5194/egusphere-2023-3111).

942 Brier, G.W., 1950. Verification of forecasts expressed in terms of probability.

943 Monthly Weather Review 78.

944 Buizza, R., 2008. The value of probabilistic prediction. Atmospheric Science

945 Letters 9, 36–42. doi:[10.1002/asl.170](http://dx.doi.org/10.1002/asl.170).

946 Burek, P., van der Knijff, J., de Roo, A., 2013. Lisflood. distributed water

947 balance and flood simulation model. doi:[10.2788/24719](http://dx.doi.org/10.2788/24719).

948 Cloke, H.L., Pappenberger, F., 2009. Ensemble flood forecasting: A review.

949 Journal of Hydrology 375, 613–626. doi:[10.1016/j.jhydrol.2009.06.](http://dx.doi.org/10.1016/j.jhydrol.2009.06.005)

950 [005](http://dx.doi.org/10.1016/j.jhydrol.2009.06.005).

951 Demeritt, D., Nobert, S., Cloke, H.L., Pappenberger, F., 2013. The european 952 flood alert system and the communication, perception, and use of ensemble 953 predictions for operational flood risk management. Hydrological Processes 954 27, 147–157. doi:[10.1002/hyp.9419](http://dx.doi.org/10.1002/hyp.9419).

955 Diomede, T., Davolio, S., Marsigli, C., Miglietta, M.M., Moscatello, A., Pa- 956 petti, P., Paccagnella, T., Buzzi, A., Malguzzi, P., 2008. Discharge pre- 957 diction based on multi-model precipitation forecasts. Meteorology and 958 Atmospheric Physics 101, 245–265. doi:[10.1007/s00703-007-0285-0](http://dx.doi.org/10.1007/s00703-007-0285-0).

959 EEA, 2023. Economic losses from weather- and climate-related ex- 960 tremes in Europe. Technical Report. European Environmental Agency. 961 URL: [https://www.eea.europa.eu/en/analysis/indicators/](https://www.eea.europa.eu/en/analysis/indicators/economic-losses-from-climate-related?activeAccordion=ecdb3bcf-bbe9-4978-b5cf-0b136399d9f8)

962 [economic-losses-from-climate-related?activeAccordion=](https://www.eea.europa.eu/en/analysis/indicators/economic-losses-from-climate-related?activeAccordion=ecdb3bcf-bbe9-4978-b5cf-0b136399d9f8)

963 [ecdb3bcf-bbe9-4978-b5cf-0b136399d9f8](https://www.eea.europa.eu/en/analysis/indicators/economic-losses-from-climate-related?activeAccordion=ecdb3bcf-bbe9-4978-b5cf-0b136399d9f8).

964 Egerton, P., Stander, J., Paterson, L., Devillier, R., Champagne-Fall, A., 965 Honore, C., Allis, E., Love, G., Nitu, R., Stuart, L., Kaya, F., Santamaria, 966 L., Teruggi, G., Taalas, P., Rea, A., Lucio, F., Harding, J., Repnik, M., 967 Luterbacher, J., Alexieva, A., Gonsior, H., Uhlenbrook, S., Cullmann, J., 968 Lombardi, N., Dunja, D., Pfister, D., Zommers, Z., Escobar Rodriguez, 969 H., Jacob, A., Keo, K., 2023. Early warnings for all. Executive action plan

970 2023-2027. Technical Report. Unitated Nations. URL: [https://www.un.](https://www.un.org/en/climatechange/early-warnings-for-all)

971 [org/en/climatechange/early-warnings-for-all](https://www.un.org/en/climatechange/early-warnings-for-all).

972 Gupta, H.V., Kling, H., Yilmaz, K.K., Martinez, G.F., 2009. Decompo- 973 sition of the mean squared error and nse performance criteria: Implica- 974 tions for improving hydrological modelling. Journal of Hydrology 377, 975 80–91. URL: [https://www.sciencedirect.com/science/article/pii/](https://www.sciencedirect.com/science/article/pii/S0022169409004843)

976 [S0022169409004843](https://www.sciencedirect.com/science/article/pii/S0022169409004843), doi:[10.1016/J.JHYDROL.2009.08.003](http://dx.doi.org/10.1016/J.JHYDROL.2009.08.003).

977 Hanssen, A., Kuipers, W., 1965. On the Relationship Between the Frequency 978 of Rain and Various Meteorological Parameters: (with Reference to the 979 Problem Ob Objective Forecasting). Koninkl. Nederlands Meterologisch 980 Institut. Mededelingen en Verhandelingen, Staatsdrukkerij- en Uitgeveri- 981 jbedrijf.

982 Joint Research Centre - European Commission, 2020. European Flood 983 Awareness System v4.0. URL: [https://confluence.ecmwf.int/](https://confluence.ecmwf.int/display/CEMS/EFAS%2Bv4.0) 984 [display/CEMS/EFAS+v4.0](https://confluence.ecmwf.int/display/CEMS/EFAS%2Bv4.0).

985 Joint Research Centre - European Commission, 2023. European Flood 986 Awareness System v5.0. URL: [https://confluence.ecmwf.int/](https://confluence.ecmwf.int/display/CEMS/Latest%2Boperational%2BEFAS%2Brelease) 987 [display/CEMS/Latest+operational+EFAS+release](https://confluence.ecmwf.int/display/CEMS/Latest%2Boperational%2BEFAS%2Brelease).

988 Kling, H., Fuchs, M., Paulin, M., 2012. Runoff conditions in the upper 989 Danube basin under an ensemble of climate change scenarios. Journal of 990 Hydrology 424-425, 264–277. doi:[10.1016/j.jhydrol.2012.01.011](http://dx.doi.org/10.1016/j.jhydrol.2012.01.011).

991 Knoben, W.J., Freer, J.E., Woods, R.A., 2019. Technical note: Inherent 992 benchmark or not? comparing nash-sutcliffe and kling-gupta efficiency 993 scores. Hydrology and Earth System Sciences 23, 4323–4331. doi:[10.](http://dx.doi.org/10.5194/hess-23-4323-2019) 994 [5194/hess-23-4323-2019](http://dx.doi.org/10.5194/hess-23-4323-2019).

995 LeClerc, J., Joslyn, S., 2015. The cry wolf effect and weather-related decision

996 making. Risk Analysis 35, 385–395. doi:[10.1111/risa.12336](http://dx.doi.org/10.1111/risa.12336).

997 Legg, T.P., Mylne, K.R., 2004. Early Warnings of Severe Weather from

998 Ensemble Forecast Information. Weather and Forecasting 19, 891–906.

999

1000

1001

Martina Angela Caretta, Aditi Mukherji, Md Arfanuzzaman, Richard

* 1. Betts, Alexander Gelfan, Yukiko Hirabayashi, Tabea Katharina Lissner, Elena Lopez Gunn, Junguo Liu, Ruth Morgan, Sixbert

1002

Mwanga, Seree Supratid, 2023. Water, in: Climate Change

|  |  |
| --- | --- |
| 1003 | 2022 – Impacts, Adaptation and Vulnerability. Cambridge University |
| 1004 | Press, pp. 551–712. URL: [https://www.cambridge.org/core/product/](https://www.cambridge.org/core/product/identifier/9781009325844%7B%25%7D23c4/type/book%7B_%7Dpart) |
| 1005 | [identifier/9781009325844{%}23c4/type/book{\_}part](https://www.cambridge.org/core/product/identifier/9781009325844%7B%25%7D23c4/type/book%7B_%7Dpart), doi:[10.1017/](http://dx.doi.org/10.1017/9781009325844.006) |
| 1006 | [9781009325844.006](http://dx.doi.org/10.1017/9781009325844.006). |
| 1007 | Molteni, F., Buizza, R., Palmer, T.N., Petroliagis, T., 1996. The ECMWF |
| 1008 | Ensemble Prediction System: Methodology and validation. Quarterly |
| 1009 | Journal of the Royal Meteorological Society 122, 73–119. doi:https: |
| 1010 | //doi.org/10.1002/qj.49712252905. |
| 1011 | Nevo, S., Morin, E., Gerzi Rosenthal, A., Metzger, A., Barshai, C., Weitzner, |
| 1012 | D., Voloshin, D., Kratzert, F., Elidan, G., Dror, G., Begelman, G., Near- |
| 1013 | ing, G., Shalev, G., Noga, H., Shavitt, I., Yuklea, L., Royz, M., Giladi, |
| 1014 | N., Peled Levi, N., Reich, O., Gilon, O., Maor, R., Timnat, S., Shechter, |
| 1015 | T., Anisimov, V., Gigi, Y., Levin, Y., Moshe, Z., Ben-Haim, Z., Hassidim, |
| 1016 | A., Matias, Y., 2022. Flood forecasting with machine learning models |
| 1017 | in an operational framework. Hydrology and Earth System Sciences 26, |
| 1018 | 4013–4032. doi:[10.5194/hess-26-4013-2022](http://dx.doi.org/10.5194/hess-26-4013-2022), [arXiv:2111.02780](http://arxiv.org/abs/2111.02780). |
| 1019 | Pappenberger, F., Cloke, H.L., Parker, D.J., Wetterhall, F., Richardson, |
| 1020 | D.S., Thielen, J., 2015. The monetary benefit of early flood warnings in |
| 1021 | Europe. Environmental Science and Policy 51, 278–291. doi:[10.1016/j.](http://dx.doi.org/10.1016/j.envsci.2015.04.016) |
| 1022 | [envsci.2015.04.016](http://dx.doi.org/10.1016/j.envsci.2015.04.016). |
| 1023 | Paprotny, D., Sebastian, A., Morales-N´apoles, O., Jonkman, S.N., 2018. |
| 1024 | Trends in flood losses in Europe over the past 150 years. Nature Com- |
| 1025 | munications 9. doi:[10.1038/s41467-018-04253-1](http://dx.doi.org/10.1038/s41467-018-04253-1). |
| 1026 | Roebber, P.J., 2009. Visualizing multiple measures of forecast quality. |
| 1027 | Weather and Forecasting 24, 601–608. doi:[10.1175/2008WAF2222159.1](http://dx.doi.org/10.1175/2008WAF2222159.1). |
| 1028 | Thielen, J., Bartholmes, J., Ramos, M.H., Roo, A.D., 2009. The european |
| 1029 | flood alert system – part 1: Concept and development. Hydrol. Earth Syst. |
| 1030 | Sci 13, 125–140. URL: [www.hydrol-earth-syst-sci.net/13/125/2009/](http://www.hydrol-earth-syst-sci.net/13/125/2009/), |
| 1031 | doi:[10.5194/hess-13-125-2009](http://dx.doi.org/10.5194/hess-13-125-2009). |
| 1032 | WMO, 2021. WMO Atlas of mortality and economic losses from weather, |
| 1033 | climate and water extremes (1970-2019). Technical Report. World Meteo- |
| 1034 | rological Organization. Geneva. |

1035

1036

1037

1038

Yang, S.C., Yang, T.H., Chang, Y.C., Chen, C.H., Lin, M.Y., Ho, J.Y., Lee,

K.T., 2020. Development of a hydrological ensemble prediction system to assist with decision-making for floods during typhoons. Sustainability (Switzerland) 12. doi:[10.3390/su12104258](http://dx.doi.org/10.3390/su12104258).