Natural Hazards
and Earth System
Sciences

Open Access

EGU

Evaluation of two hydrometeorological ensemble strategies for flash-flood forecasting over a catchment of the eastern Pyrenees

Hélène Roux[1], Arnau Amengual[2], Romu Romero[2], Ernest Bladé[3], and Marcos Sanz-Ramos[3]

[1]Institut de Mécanique des Fluides de Toulouse (IMFT), Université de Toulouse, CNRS, Toulouse, France
[2]Grup de Meteorologia, Departament de Física, Universitat de les Illes Balears, Palma, Majorca, Spain
[3]Institut FLUMEN, E.T.S. d'Eng. De Camins, Canals i Ports de Barcelona,
Universitat Politècnica de Catalunya, Barcelona, Spain

Correspondence: Hélène Roux (helene.roux@imft.fr)

Abstract. This study aims at evaluating the performances of flash-flood forecasts issued from deterministic and ensemble meteorological prognostic systems. The hydrometeorological modeling chain includes the Weather Research and Forecasting Model (WRF) forcing the rainfall-runoff model MARINE dedicated to flash floods. Two distinct ensemble prediction systems accounting for (i) perturbed initial and lateral boundary conditions of the meteorological state and (ii) mesoscale model physical parameterizations have been implemented on the Agly catchment of the eastern Pyrenees with three subcatchments exhibiting different rainfall regimes.

Different evaluations of the performance of the hydrometeorological strategies have been performed: (i) verification of short-range ensemble prediction systems and corresponding streamflow forecasts, for a better understanding of how forecasts behave; (ii) usual measures derived from a contingency table approach, to test an alert threshold exceedance; and (iii) overall evaluation of the hydrometeorological chain using the continuous rank probability score, for a general quantification of the ensemble performances.

Results show that the overall discharge forecast is improved by both ensemble strategies with respect to the deterministic forecast. Threshold exceedance detections for flood warning also benefit from large hydrometeorological ensemble spread. There are no substantial differences between both ensemble strategies on these test cases in terms of both the issuance of flood warnings and the overall performances, suggesting that both sources of external-scale uncertainty are important to take into account.

1 Introduction

Flash floods are among the most devastating natural hazards worldwide, producing important human and socioeconomic losses. The western Mediterranean region is annually affected by several extreme precipitation events which lead to flash flooding. During the extended warm season, the early intrusion of upper-level cold air masses and the relatively high sea surface temperature boost the convective available potential energy of the low-level Mediterranean warm and moist air. This natural hazard results from the persistence of deep moist convection and intense precipitation over specific hydrographic catchments during several hours. As many western Mediterranean small-to-medium-sized river basins are highly urbanized, steep and close to the coastline, their hydrological responses are inherently short. Large, rapid and unexpected flows exacerbate flood damage. The development and evaluation of the state-of-the-art hydrometeorological forecasting tools is a major issue in the Hydrological cycle in the Mediterranean experiment (HyMeX; Drobinski et al., 2014). This program aims at addressing the following science questions, amongst others. How can we improve heavy rainfall process knowledge and prediction? How can we improve hydrological prediction?

Hydrometeorological forecasting tools can contribute to a better understanding and forecasting of flash floods so as to implement more reliable forecasting and warning systems over the western Mediterranean. Short-range quantitative precipitation forecasts (QPFs) by high-resolution numerical weather prediction (NWP) models are an effective

tool to further extend flood forecasting lead times beyond the basin response times. NWP models capture the initiation and evolution of small-scale and convectively driven precipitations, with similar spatial and temporal scales to the flash-flood-prone catchments (Leoncini et al., 2013; Fiori et al., 2014; Ravazzani et al., 2016; Amengual et al., 2017). Although QPFs can be directly used to force one-way hydrological models, the hydrometeorological forecasts are impacted by different types of uncertainties. Uncertainties are inherent to each of the hydrometeorological chain components: model parameterization and structure, limitations of measuring devices providing observation data, and initial and lateral boundary conditions (Zappa et al., 2010).

External-scale inaccuracies in the hydrological models emerge from two distinct sources when forecasting deep moist convection and heavy rainfall with NWP models. First, errors arise from the complexity and nonlinearity of the physical parameterizations. Second, uncertainties emerge when representing the exact initial atmospheric state and boundary forcing across the scales where convection develops. But reliable spatial and temporal QPF distributions are necessary to render skillful quantitative discharge forecasts when coping with floods over small- and medium-sized basins. Otherwise, the issuance of precise and dependable early flood warnings is inhibited (Le Lay and Saulnier, 2007; Bartholmes et al., 2009; Cloke et al., 2013).

To alleviate the impact of these external-scale uncertainties, short-range ensemble prediction systems (SREPSs) are used to build hydrological ensemble prediction systems (HEPSs). SREPSs aim at sampling the set of plausible outcomes and at accounting for the most relevant uncertainties in the atmospheric forecasting process so as to increase. Uncertainties in the initial and boundary fields can be encompassed by conveniently perturbing initial and lateral boundary conditions (IC/LBCs, Grimit and Mass, 2007; Hsiao et al., 2013). Uncertainties in model parameterizations are dealt with by populating the ensemble with multiple combinations of equally skillful physical schemes (Stensrud et al., 2000; Jankov et al., 2005; Amengual et al., 2008, 2017; Tapiador et al., 2012). The inclusion of these uncertainties aims at improving the skill and spread of the HEPSs by introducing independent information of all the plausible atmospheric states and processes. Therefore, SREPSs are increasingly used in hydrologic prediction (Cloke and Pappenberger, 2009; Verkade et al., 2013, 2017; Siddique and Mejia, 2017; Benninga et al., 2017; Bellier et al., 2017, 2018; Edouard et al., 2018; Jain et al., 2018). Several studies have stated that probabilistic forecasts could improve decision-making if appropriately handled (e.g., Krzysztofowicz, 2001; Todini, 2004; Ramos et al., 2013; Antonetti et al, 2019). As stated by Zappa et al. (2011), each member of a meteorological ensemble can be fed into a hydrological model to generate a hydrological forecast.

However, the most appropriate methods for generating HEPSs and the quantification of their added value are still under assessment (Cloke and Pappenberger, 2009; Cloke et al., 2013). Further efforts devoted to examine the predictive skill of both ensemble strategies and how the external-scale uncertainties spread into the HEPSs become paramount for the optimal design of hydrometeorological operational chains over the flood-prone western Mediterranean area. The objective of the present work is to evaluate the predictive skill of two distinct HEPS generation strategies – accounting for perturbed IC/LBCs (PILB) and mixed physics (MPS) – for three flash-flood episodes over the Agly basin (Fig. 1). This catchment of the eastern Pyrenees has been selected as an experimental area as several subcatchments exhibit different rainfall regimes. Given the small size of the subcatchments (from 150 to 300 km^2), the localization of the precipitation patterns is crucial (Rossa et al., 2010), and it is a challenge to implement QPFs for such small subcatchments. QPFs are generated by using the Weather Research and Forecasting Model (WRF; Skamarock et al., 2008). Next, 48 h WRF forecasts are propagated down through the MARINE hydrological model (Roux et al., 2011) to investigate the quantitative discharge forecasts in the timing and magnitude of these flash floods. The resulting HEPSs are examined using different criteria to illustrate the potential benefits of the probabilistic hydrometeorological forecast chains. The rest of the paper is structured as follows: Sect. 2 presents a short overview of the flash floods, the study area and the observational networks; Sects. 3 and 4 provide an insight into the hydrological and atmospheric models and the strategies for ensemble generation; and Sect. 5 presents the results. The last section summarizes main conclusions and provides further remarks.

2　Data and case studies

2.1　Overview of the Agly catchment

This study focuses on a catchment in the north side of the eastern Pyrenees, the Agly, as a test site for implementing the HEPS strategies. The Agly is a coastal river in the north side of the eastern Pyrenees (Fig. 1). It originates from an elevation of approximately 700 m and drains the Pyrenees foothills. It flows into the Mediterranean Sea at Le Barcarès and has a length of around 80 km. A dam dedicated to flood and water management controls approximately 400 km^2 of the catchment (Agence de l'eau Rhône Méditerranée & Corse, 2012). It is located just downstream of the confluence between the Agly and one of its main right-hand tributaries, the Désix river, draining an area of around 150 km^2 (Fig. 1). The main left-hand tributary, the Verdouble river, drains an area of 300 km^2 located in a midmountain region, culminating between 400 and 500 m of altitude (Fig. 1). Granite and gneiss cover about 300 km^2 of the mountainous part of the Agly catchment, promoting runoff already facilitated by the steep slopes. North of the catchment, the Corbières massif is dominated by limestones form-
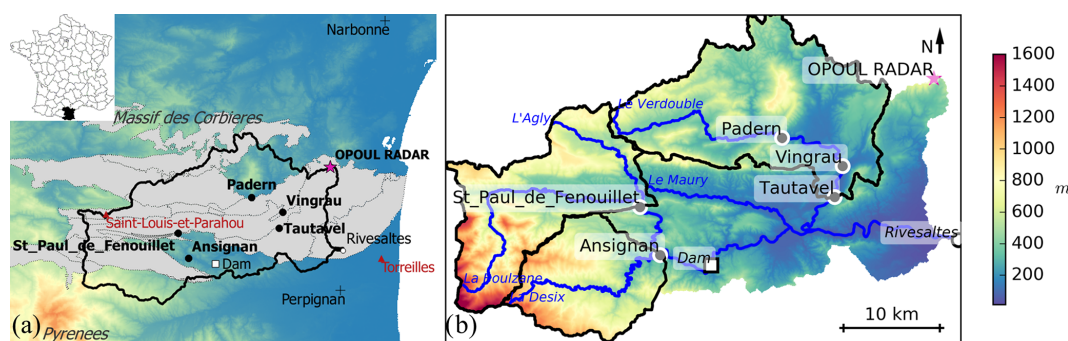
Figure 1. (a) Location of the Agly catchment and of the meteorological radar (grey area: karstic areas underlying the Agly catchment, from BDLISA version 2: Base de Donnée des Limites des Systèmes Aquifères, https://bdlisa.eaufrance.fr/, last access: 18 June 2019). **(b)** Digital terrain model of the Agly catchment (source: IGN; MNT BDALTI). Also included are the main tributaries (blue lines, source: IGN, BD CARTHAGE), the radar location (pink star: OPOUL RADAR), the discharge gaging stations (black dots), the dam (white square) and the outlet (white circle).

ing karstic networks. According to hydrogeological studies of the area, there are only losses in the Agly and Verdouble catchments due to the karstic system. These losses contribute to the streamflow of two resurgences draining the Corbières massif but located outside of the Agly catchment (Font Estramar and Font Dame resurgences) (Salvayre, 1989). The average loss rates are estimated between 0.3 and 1.5 m^3 s^{-1} for the Agly, depending on the river discharge, and between 0.7 and 2 m^3 s^{-1} on the Verdouble (Ladouche and Dörfliger, 2004). These are only average estimates based on observed discharges and assumptions about the functioning of the karst system, but they can be considered small enough not to be explicitly represented in flash-flood simulations. A total of 80 % of the catchment is covered by natural vegetation – forest (45 %), shrubby vegetation (17 %), maquis and scrubland (16 %) – while 18 % is used for agriculture, mainly vineyards.

The Agly catchment is subject to different climate regimes in connection with the distances from the sea and the mountainous reliefs: temperate oceanic in the north-west valley, mountain in the south-west part, and Mediterranean downstream. The rainfall regime varies from east to west with increasing annual cumulated precipitations: the mean annual cumulated precipitations (1965–1996) range from 600 mm at Torreilles (east, Fig. 1) up to 1174 mm at Saint-Louis-et-Parahou (west, Fig. 1) (DIREN Languedoc-Roussillon/SIEE-GINGER, 2008). Generally, the rainfall regime is highly variable, with very intense precipitation events in fall, winter, and spring and with very dry summers.

2.2 Available data

The precipitation measurements available on the Agly catchment come from two different observational networks.

– PLU: the operational hourly rain-gauge network for flood-monitoring purposes and data provided by the re-

gional flood forecasting service, the Service de Prévision des Crues Méditerranée Ouest (SPCMO);

– JP1: 1 km^2 quantitative hourly precipitation estimates ANTILOPE $J+1$ (ANalyse par spaTIaLisation hOraire des PrEcipitations) that come from a merging of radar data and rain-gauge measurements (Laurantin, 2008; Champeaux et al., 2009).

The hydrometric data were derived from the network of operational measurements at variable time steps (HydroFrance databank, http://www.hydro.eaufrance.fr/, last access: 20 November 2019). The stream gauges are located in five upstream stations not influenced by the dam (Table 1 and Fig. 1). Table 2 summarizes the main hydrological features of the five stations. This study will focus on three recent events beginning on 4 March 2013, 16 November 2013 and 28 November 2014, being highly variable (Table 3), with rainfall lasting respectively 3 d for the spring event and 4 d for the two fall events. The selected events have been labeled with the start date and the duration as follows: 20130304_3d, 20131116_4d and 20141128_4d. All the floods feature moderate specific peak discharges for flash floods, highlighting the high infiltration rates. The runoff coefficient is always higher for the eastern part (station no. 5, Table 3) than for the western part. The runoff coefficient is even higher than 1 for 20130304_3d at station no. 5. There is no definitive explanation for that, but several possibilities can be considered: (i) the very high soil moisture at the beginning of the event (65 %, Table 3), which can contribute to the runoff at the outlet via subsurface flows; (ii) an amount of snowmelt as there was a snowfall episode at the very end of February 2013 over the eastern Pyrenees and Corbières, with snow above 700 to 800 m; (iii) the uncertainties in the discharge and precipitation measurements; (iv) a possible supply from the karstic system (Fig. 1); however, this possibility is pretty unlikely as hydrological studies conclude to only losses in the Verdouble catchments due to the karstic

Table 1. Characteristics of the five subcatchments and the whole catchment. The time of concentration is estimated using the Bransby Williams formula (Eq. 3).

Station	River	Area (km^2)	T_c (h)
No. 1: Ansignan	Désix	157	9
No. 2: Saint-Paul-de-Fenouillet	Agly	216	10
No. 3: Padern	Verdouble	161	8
No. 4: Vingrau	Verdouble	301	11
No. 5: Tautavel	Verdouble	305	12
Rivesaltes	Agly	1053	23

Table 2. Hydrological statistics of the five catchments (from HydroFrance databank, http://www.hydro.eaufrance.fr/, last access: 20 November 2019). QIX2: 2-year return period of maximum instantaneous discharge and 95 % confidence interval; QMEV: known maximum instantaneous discharge; TMEV: date of QMEV.

Station	Period	QIX2 (m^3 s^{-1})	QMEV (m^3 s^{-1})	TMEV
No. 1	1994–2018	85.0 [57.00; 120.0]	291	15 Mar 2011
No. 2	1971–2018	87.0 [77.00; 99.00]	483	26 Sep 1992
No. 3	2006–2018	–	281	30 Nov 2014
No. 4	2010–2018	–	525	30 Nov 2014
No. 5	1967–2018	170.0 [140.0; 200.0]	922	13 Nov 1999

system (Ladouche and Dörfliger, 2004). One event occurred in spring with an averagely moist soil (20130304_3d, Table 3), while the other two occurred in autumn with dry soils after the summery drought. The autumn episodes exhibit very different intensities: the specific peak discharges range from 0.3 to 0.6 m^3 s^{-1} km^{-2} for 20131116_4d and from 1 to 2 m^3 s^{-1} km^{-2} for 20141128_4d. Concerning the means of the maximum rainfall intensity over the catchment, they range from 8 to 14 mm h^{-1} according to PLU and from 9 to 11 mm h^{-1} according to JP1 for 20131116_4d as well as from 19 to 30 mm h^{-1} according to PLU and from 15 to 25 mm h^{-1} according to JP1 for 20141128_4d (Table 3). 20141128_4d is therefore much more intense than 20131116_4d according to both observed forcings even if JP1 forcing presents lower intensities. 20130304_3d is in between both episodes, with specific peak discharges ranging from 0.6 to 1.5 m^3 s^{-1} km^{-2} but lower rainfall intensities ranging from 7 to 11 mm h^{-1} according to PLU and from 6 to 11 mm h^{-1} according to JP1. These episodes are representative of the different seasonal rainfall regimes that lead to floods over the Agly. In spring, floods mainly originate from stratiform-type rainfall with moderate but persistent precipitation rates that can result in substantial accumulations. In autumn, floods are most likely driven by convective-type precipitations of shorter duration but high intensity.

Figure 2 shows the spatial repartition of the cumulative rainfall for the three events for both forcings. The rain-

gauge data have been interpolated using the Thiessen polygon methods (Thiessen, 1911). Variability in rainfall clearly emerges especially between the eastern, western and mountainous part.

3   Hydrological tool

3.1   Rainfall-runoff model

The MARINE model is a distributed mechanistic hydrological model specially developed for flash-flood simulations. It models the main physical processes in flash flooding: infiltration, overland flow, and lateral flows in soil and channel routing. Conversely, it does not incorporate low-rate flow processes such as evapotranspiration or base flow.

MARINE is structured into three main modules that are run for each catchment grid cell (Fig. 3). The first module allows the separation of surface runoff and infiltration using the Green–Ampt model (Green and Ampt, 1911). The second module represents subsurface downhill flow, based on the generalized Darcy law used in the TOPMODEL hydrological model (Beven and Kirkby, 1979). Lastly, the third module represents overland and channel flows. Rainfall excess is transferred to the catchment outlet using the Saint-Venant equations simplified with kinematic wave assumptions (Fread, 1992). The model distinguishes grid cells with a drainage network, where channel flow is calculated on a triangular channel section (Maubourguet et al., 2007); and from grid cells on hillslopes, where overland flow is calculated for the entire surface area of the cell. For more details about the MARINE model, the readers can refer to Roux et al. (2011), Garambois et al. (2015b) and Douinot et al. (2018).

The MARINE model works with distributed input data such as (i) a digital elevation model (DEM) of the catchment to shape the flow pathway and distinguish hillslope cells from drainage network cells, according to a drained area threshold; (ii) soil survey data to initialize the hydraulic and storage properties of the soil, which are used as parameters in the infiltration and lateral flow models; and (iii) vegetation and land-use data to configure the surface roughness parameters used in the overland flow model. As the MARINE model is event-based, it must be initialized to take into account the previous moisture state of the catchment. This is done by using the spatial daily root-zone saturation state, i.e., the ratio of the soil water content to the soil storage capacity at a spatial resolution of 8 km × 8 km, output from Météo France's SIM operational chain (Habets et al., 2008). The initial soil water content for MARINE is therefore directly obtained by multiplying the saturation state by the soil storage capacity of each cell.

3.2   Calibration/validation on the Agly catchment

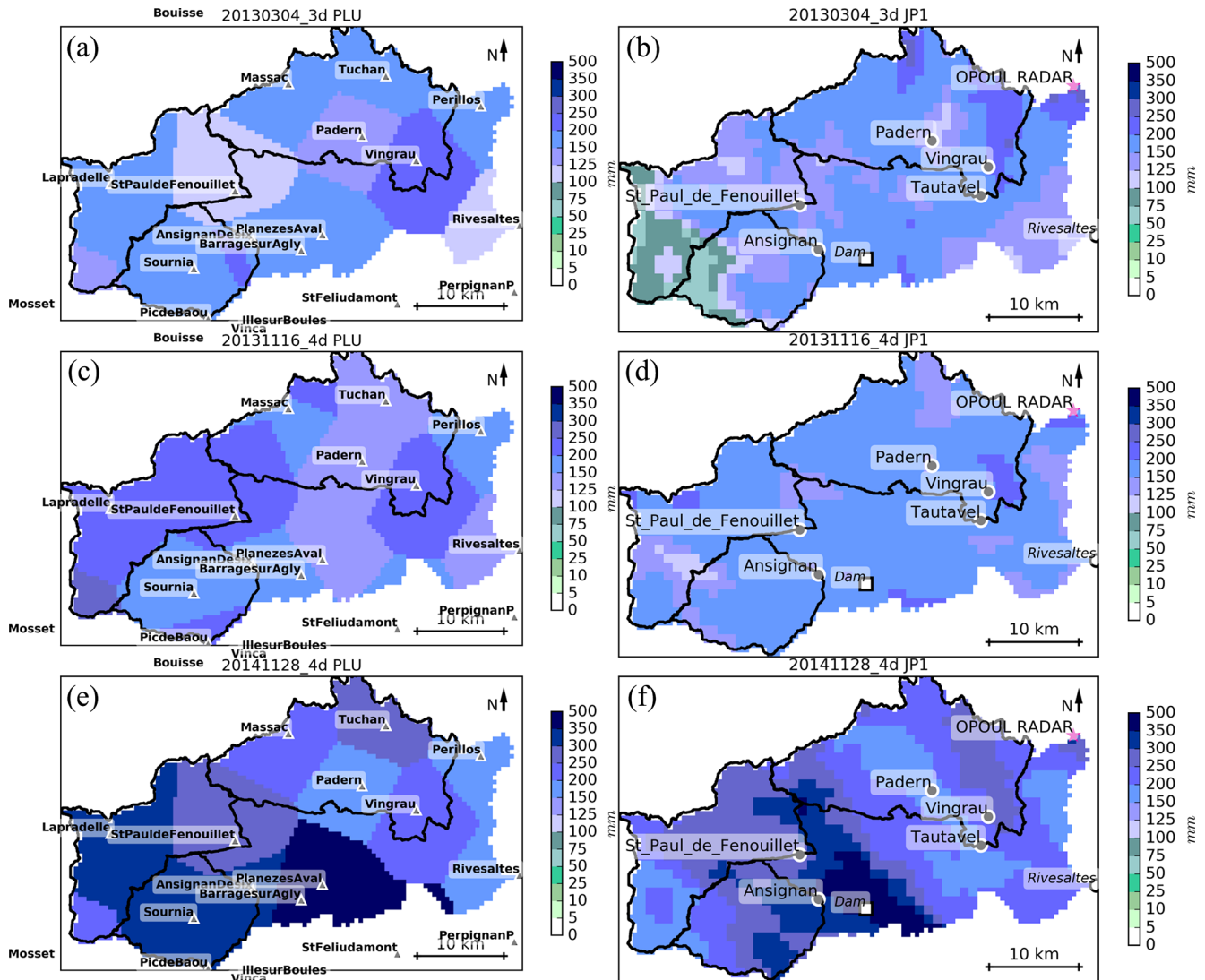MARINE requires parameter calibration so as to accurately reproduce hydrological behaviors. Based on previous sen-

Figure 2. Spatial variability of the cumulative rainfall for event 20130304_3d (**a, b**), 20131116_4d (**c, d**) and 20141128_4d (**e, f**), according to the observations. PLU (**a, c, e**): the operational hourly rain-gauge network (from Hydroreel, Serveur de données hydrométriques en temps réel, Bassin Rhône-Méditerranée et Région Auvergne-Rhône-Alpes, https://www.rdbrmc.com/hydroreel2/listestation.php, last access: 20 November 2019); JP1 (**b, d, f**): 1 km^2 merging of radar data and rain-gauge measurements.

sitivity analyses by Garambois et al. (2013), five parameters are calibrated: soil depth C_Z; the transmissivity used in lateral subsurface flow modeling C_T; hydraulic conductivity at saturation C_K; and friction coefficients for low- and high-water channels, n_L and n_H, respectively. C_T, C_K and C_Z are the multiplier coefficients for spatialized, saturated hydraulic conductivities and soil depths. Note that n_L and n_H are kept invariant throughout the drainage network. The spatial resolution of the MARINE model on all the Agly subcatchments is of 500 m. The calibration of the Agly catchment at the Saint-Paul-de-Fenouillet station (no. 2, Table 1 and Fig. 1) was performed by Garambois et al. (2015a) according to their proposed methodology. The events used for this calibration are older than those considered in the

present study (20020411, 20031204, 20040221, 20051115, 20101010, 20110315; see Garambois et al., 2015a). The cost function L_{NP} is designed to evaluate the performance of the model (Roux et al., 2011; Garambois et al., 2015a):

$$L_{NP} = \frac{1}{3}L_N + \frac{1}{3}\left(1 - \frac{\left|Q_p^s - Q_p^o\right|}{Q_p^o}\right) + \frac{1}{3}\left(1 - \frac{\left|T_p^s - T_p^o\right|}{T_c}\right), \quad (1)$$

where Q_p^s and Q_p^o are respectively the simulated and observed peak runoff, T_p^s and T_p^o are the simulated and observed time to peak, and T_c is the time of concentration of the catchment. L_N denotes the efficiency coefficient (Nash and Sutcliffe, 1970):

Table 3. Main features of the selected flash-flood events. Observed forcing PLU: network of 19 rain gauges; observed forcing JP1: 1 km^2 quantitative precipitation estimates; cumulated P (mm): mean $\pm$ standard deviation [max] of accumulated precipitation on the catchment during the whole event; max I (mm h^{-1}): mean of the maximal rainfall intensity over the catchment; Q_p^o (m^3 s^{-1}): peak discharge for the event; $Q_p^o\big|_s$ (m^3 s^{-1} km^{-2}): ratio of the peak discharge for the event to the drainage area of the subcatchment; T_p^o (dd hh:mm): date of the peak discharge; C_r (–): observed runoff coefficient – ratio of the amount of runoff through the outlet to the amount of rainfall on the catchment; H_{ini} (%): mean $\pm$ standard deviation initial soil moisture according to Safran-Isba-Modcou (SIM) daily root-zone humidity output (Habets et al., 2008).

| Event | Station | PLU | | JP1 | | Q_p^o | $Q_p^o\big|_s$ | T_p^o | C_r | H_{ini} |
| | | Cumulated P | Max I | Cumulated P | Max I | | | | | |
		(mm)	(mm h^{-1})	(mm)	(mm h^{-1})	(m^3 s^{-1})	(m^3 s^{-1} km^{-2})	(dd hh:mm)		(%)
20130304_3d	no. 1	186 ± 19 [226]	7.4	167 ± 30 [208]	6.4	137	0.87	06 06:35	0.17	48 ± 0
	no. 2	183 ± 37 [215]	6.9	160 ± 25 [217]	5.8	137	0.63	06 09:40	0.12	51 ± 3
	no. 5	181 ± 28 [218]	11.2	192 ± 26 [294]	11.4	459	1.50	06 12:24	1.07	65 ± 1
	outlet	179 ± 40 [226]	8.5	178 ± 30 [294]	8.6	970	0.92	–	–	56 ± 7
20131116_4d	no. 1	227 ± 11 [303]	13.1	208 ± 18 [242]	10.9	47	0.30	18 05:10	0.05	35 ± 1
	no. 2	275 ± 26 [303]	14.1	212 ± 24 [269]	8.8	131	0.61	18 01:58	0.05	42 ± 4
	no. 5	181 ± 37 [241]	8.0	183 ± 17 [230]	10.6	109	0.36	18 06:13	0.21	55 ± 3
	outlet	208 ± 49 [303]	9.9	194 ± 25 [285]	9.6	260	0.25	–	–	45 ± 8
20141128_4d	no. 1	311 ± 12 [318]	30.4	284 ± 40 [361]	25.0	251	1.60	30 14:56	0.14	36 ± 0
	no. 2	286 ± 28 [312]	18.8	261 ± 41 [357]	15.1	215	0.99	29 22:28	0.07	40 ± 4
	no. 5	222 ± 37 [264]	20.9	234 ± 36 [356]	20.7	606	1.99	30 07:45	0.67	58 ± 5
	outlet	269 ± 61 [392]	14.5	257 ± 54 [492]	12.8	978	0.93	–	–	48 ± 10

Figure 3. Structure of the MARINE model.

$$L_N = 1 - \frac{\sum_{i=1}^{n}\left(Q_i^s - Q_i^o\right)^2}{\sum_{i=1}^{n}\left(Q_i^o - \overline{Q^o}\right)^2}, \tag{2}$$

where n is the number of observation data, and Q^s and Q^o are the simulated and the observed runoff. The estimated times of concentration of each subcatchment are given in Table 1, using the Bransby Williams formula (Pilgrim and Cordery, 1992):

$$T_c = 14.6 L A^{-0.1} S^{-0.2}, \tag{3}$$

where T_c (min) is the time of concentration, L (km) is the total length of the channel, A (km^2) is the drainage basin area

and S (m m^{-1}) is the average slope. Here, the formula for time of concentration is only used to normalize the peak time delay in the third term of Eq. (1) with a characteristic time of the catchment, so the most important point is to always use the same procedure to make this term dimensionless. Note that the range of values for both L_{NP} and L_N spans from $-\infty$ to 1, with 1 being the perfect score.

Table 4 lists the L_N and L_{NP} efficiencies for the validation cases: the three studied events with different forcings and two older flash-flood events with available data, only used for the validation process of the hydrological model but not further studied. Table 4 and Fig. 4 show the following.

– Only one event (20130304_3d with PLU forcing) is well simulated at the five gauging stations.

– Only one event (20130304_3d with both PLU and JP1 forcings) is well simulated at mountainous station no. 1.

– All the other events are correctly simulated only for a part of the catchment: either the eastern part near the Mediterranean Sea (stations no. 3, no. 4 and no. 5), the south-west mountainous part (station no. 1), or the north-west continental part (station no. 2). This result does not seem to be directly linked with the rain-gauged distribution because, first of all, the rain-gauge network is quite dense in this catchment and rather well distributed: with 19 rain gauges for an area of around 1000 km^2, the rain-gauge density is about 1 for 50 km^2, whereas the rain-gauge density for the full network over mainland France is of 1 for 120 km^2 (Mounier et al.,
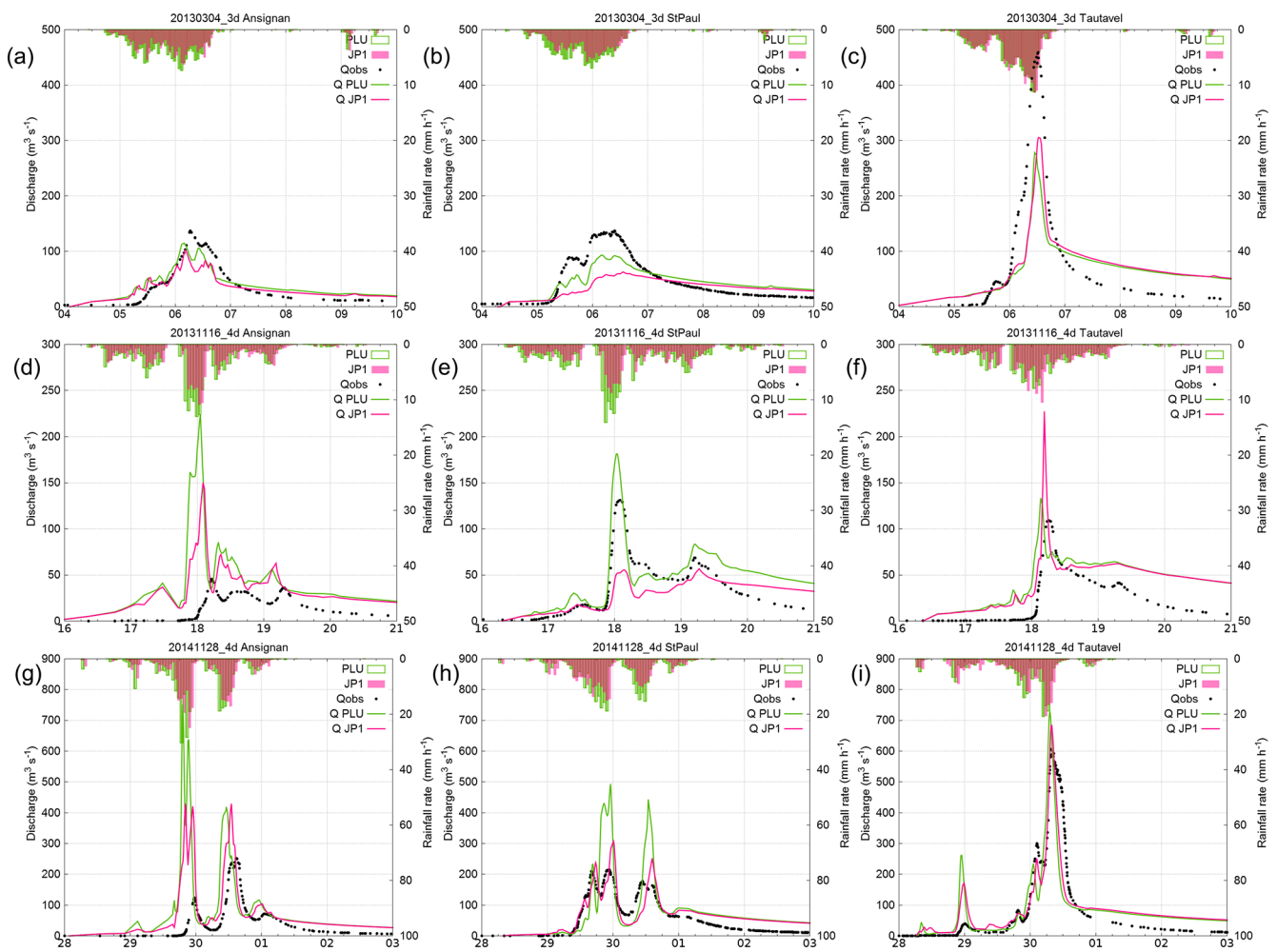
Figure 4. Hyetograph and hydrograph at station no. 1 (**a, d, g**), no. 2 (**b, e, h**) and no. 5 (**c, f, i**) for three events. PLU: forcing with the network of 19 rain gauges; JP1: forcing with 1 km^2 quantitative precipitation estimates; Q_{obs}: observed discharge at the station; Q_{PLU}: simulated discharge with PLU forcing; Q_{JP1}: simulated discharge with JP1 forcing.

Table 4. $L_{NP}(L_N)$ efficiencies for each station (see numbering in Table 1) and for each validation event. PLU: forcing with the network of 19 rain gauges; JP1: forcing with 1 km^2 quantitative precipitation estimates. Bold values indicate efficiencies above 0.5.

Event forcing	No. 1	No. 2	No. 3	No. 4	No. 5
19920926_PLU	–	**0.92 (0.93)**	–	–	
20090411_PLU	< 0 (< 0)	**0.50** (0.12)	< 0 (< 0)	–	< 0 (< 0)
20130304_3d_PLU	**0.78 (0.80)**	**0.61 (0.72)**	**0.61** (0.43)	**0.67 (0.60)**	**0.70 (0.61)**
20130304_3d_JP1	**0.74 (0.73)**	< 0 (0.34)	**0.67 (0.52)**	**0.77 (0.66)**	**0.78 (0.69)**
20131116_4d_PLU	< 0 (< 0)	**0.64** (0.41)	0.06 (< 0)	< 0 (< 0)	0.38 (< 0)
20131116_4d_JP1	< 0 (< 0)	< 0 (0.36)	< 0 (< 0)	< 0 (< 0)	0.24 (< 0)
20141128_4d_PLU	< 0 (< 0)	0.11 (< 0)	**0.65** (0.16)	**0.67** (0.47)	**0.79 (0.61)**
20141128_4d_JP1	< 0 (< 0)	**0.68 (0.64)**	**0.78 (0.73)**	**0.81 (0.74)**	**0.89 (0.81)**

2012). In addition, it is not always for the same part of the catchment that the model has the best performance: it depends on the event. Therefore, the same distribution of rain gauges sometimes leads to a correct simulation in terms of L_{NP} cost function (Eq. 1) for a given even,

while it leads to an unsatisfactory simulation for another event.

As expected, the different parts of the catchment exhibit various behaviors which are difficult to correctly simulate with a single calibration by just using observations at sta-

tion no. 2. On the one hand, events with relatively moderate peak discharge are usually not correctly simulated by MARINE regardless of the observed forcing, as is the case of the 20090411_PLU and 20131116_4d events. Indeed, several authors have pointed out that specific peak discharges larger than $0.5\,\mathrm{m^3\,s^{-1}\,km^{-2}}$ are one of the relevant criteria to define a flash flood (Braud et al., 2014; Gaume et al., 2009). The 20090411_PLU and 20131116_4d events exhibit smaller peak discharges (Table 3), except for the 20131116_4d episode at station no. 2, where the results are correct for the PLU forcing (Fig. 4). When the simulated hydrographs are suitable for the eastern Agly, the discharge is overestimated over the western part (e.g., 20141128_4d; Fig. 4). Conversely, when the simulated hydrographs are correct over the western Agly, the peak discharges are underestimated in the eastern part as in the 20130304_3d episode. Difficulties in correctly simulating the hydrological responses over all the subcatchments arise due to the spatial variability of hydrological behavior across the Agly catchment, leading to a myriad of runoff responses that are difficult to encompass with single parameterizations of the infiltration process in hydrological models (Amengual et al., 2017).

With respect to the two major 20130304_3d and 20141128_4d events, both simulated with the two observed forcings, simulations are more satisfactory with the $1\,\mathrm{km^2}$ quantitative precipitation estimates ANTILOPE $J+1$ for the eastern than for the western part. This may be due to the fact that the radar is located close to the sea, with the beams being orographically sheltered over the western Agly (Fig. 1). Several other calibration tests could have been carried out so as to improve the results of the hydrological model such as one calibration for each subcatchment. However, the main purpose of this study focuses on the potential of ensemble strategies to improve flash-flood forecasting. Furthermore, NWP-model-driven runoff simulations have been compared both against the observed discharges and against the observed rain-gauge and radar-precipitation-driven runoff runs. Hence, the impact of the external-scale uncertainties on the quality of the distinct HEPS can be emphasized.

4 Meteorological tools

The fully compressible and nonhydrostatic WRF model has been employed to generate the ensemble members. The WRF setup consists of a single computational domain completely spanning the western Mediterranean region at 2.5 km spatial horizontal resolution (i.e., 767×575 grid points) and 50 vertical levels (Fig. 5). Deep moist convection is explicitly solved due to the high-spatial resolution. All the ensemble experiments have a temporal forecasting horizon of 48 h, starting at 00:00 UTC on the day before of the main observed peak floods. Starting on this day guarantees a suitable lead time to issue warnings to local water management services. For these hydrometeorological episodes lasting more than 2 d, successive consecutive 48 h simulations have been performed, starting on the next days at 00:00 UTC. Hence, the initiation and subsequent evolution of the most active precipitation systems and the overall rainfall episodes are completely encompassed.

WRF simulations have been forced by using the global Ensemble Prediction System of the European Centre for Medium-Range Weather Forecasts (ECMWF-EPS). The MPS ensemble has been built by using the reference (i.e., unperturbed) run, while the PILB approach has considered a selected set of the overall ECMWF-EPS population. Finally, the hourly QPFs are used to force the MARINE model one way so as to build the HEPSs. In addition, the deterministic ECMWF forecasts have been also dynamically downscaled so as to have a control baseline for comparative purposes against the ensemble strategies.

Deterministic simulations have used the following physical parameterizations: the WRF single-moment six-class microphysics scheme, including graupel (WSM6; Hong and Lim, 2006); the 1.5-order Mellor–Yamada–Janjić boundary layer scheme (MYJ; Janjić, 1994); the Dudhia shortwave scheme (Dudhia, 1989); the RRTM longwave scheme (Mlawer et al., 1997); the unified Noah land surface model (Tewari et al., 2004); and the Eta similarity surface-layer model (Janjić, 1994). Note that the WRF configuration for the control simulations is the same as the daily operational setup run by the research Meteorology Group at the University of the Balearic Islands (http://meteo.uib.es/wrf, last access: 31 January 2020).

4.1 PILB ensemble

The operational ECMWF-EPS is formed by 51 members – the reference and 50 perturbed forecasts – at T639 spectral resolution (20 km) and aims to cope with uncertainties related to the actual state of the atmosphere. The daily synoptic-scale uncertainties are encompassed by perturbing an initial analysis through the flow-dependent singular vector technique (Buizza and Palmer, 1995; Molteni et al., 1996). However, perturbed IC/LBCs can produce inadequate spread in the short range, before error growth on the synoptic scale becomes nonlinear (Gilmour et al., 2001). Therefore, the implemented PILB ensemble is based on dynamically downscaling these 20 ECMWF-EPS members exhibiting maximum perturbations in the initial and lateral boundary conditions over the WRF domain. This strategy seeks to ameliorate the aforementioned mismatch between the synoptic-scale error growth optimization time for the singular vectors and the subsynoptic error growth, which is more relevant for short-range forecasts at small- and medium-sized basins (Ravazzani et al., 2016; Amengual et al., 2017).

At this aim, a k-means clustering algorithm using the principal components of the 500 hPa geopotential and 850 hPa temperature fields is applied to the entire ECMWF-EPS over the WRF numerical domain. Then, the 50 ensemble members
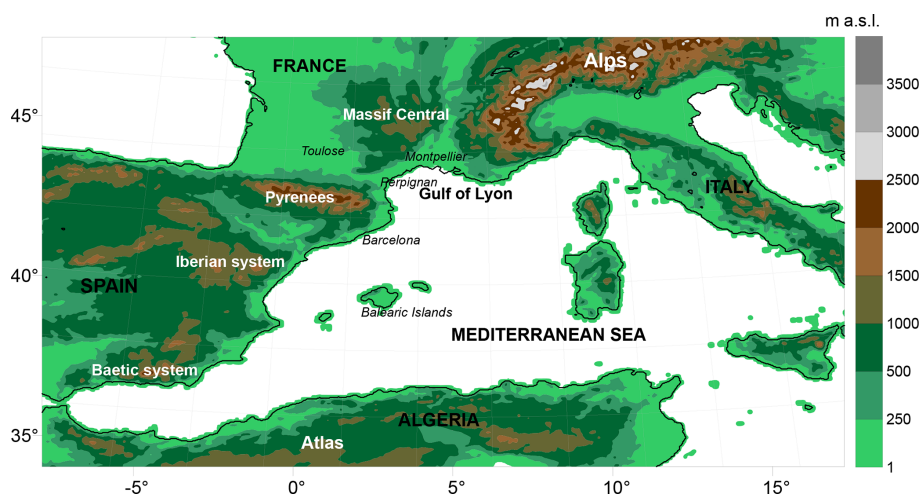
Figure 5. Configuration of the computational domain used for the WRF numerical simulations.

are categorized in 20 clusters and the 20 closest members to the centroids are used as initial and boundary fields for the PILB ensemble. Boundary fields are updated every 3 h, and physical schemes remain invariant for all the ensemble members and are the same as those used to run the deterministic WRF simulations.

4.2 Mixed-physics (MPS) ensemble

There is not an optimum set of physical numerical parameterizations when simulating severe weather and intense precipitation events. Several studies have shown that different combinations of physical parameterizations render similar performances (Jankov et al., 2005; Evans et al., 2012). That is, the meteorological variables are sensitive to a myriad of processes which are differently parameterized by capable numerical schemes. When simulating flash flooding driven by convective-type precipitation, cumulus parameterizations are the main candidates for direct uncertainty sampling. However, as convection is explicitly resolved, uncertainties arising from the microphysical subgrid processes and planetary boundary layer (PBL) schemes have been encompassed. The former regulates the distinct forms of rainfall; the latter accounts for the turbulent vertical fluxes of heat, momentum, and moisture within the PBL and throughout the atmosphere. Both physical mechanisms are also dominant when controlling deep moist convection. The MPS ensemble has been generated using all possible pairs (cloud microphysics-boundary layer) between the following schemes, summing up to 20 members.

- Microphysical schemes: (i) WRF single-moment six-class (WSM6; Hong and Lim, 2006); (ii) Goddard (Tao et al., 1989); (iii) New Thompson (Thompson et al., 2008); and (iv, v) National Severe Storms Laboratory (NSSL) two-moment (Mansell, 2010) with two

cloud condensation nuclei (CCN) prediction values of 0.5×10^9 and 1.0×10^9 cm^{-3}.

- PBL schemes: (i) Yonsei University (YSU; Hong et al., 2006); (ii) Mellor–Yamada–Janjić (MYJ; Janjic, 1994); (iii) Mellor–Yamada–Nakanishi–Niino level 2.5 (MYNN; Nakanishi and Niino, 2006), and (iv) total energy–mass flux (TEMF; Angevine et al., 2010).

On the one hand, all microphysics schemes involve the simulation of explicitly resolved liquid water, cloud, and precipitation and include mixed-phase transformations (i.e., the interaction of ice and liquid water). However, each microphysical parameterization treats differently the interaction among five or six moisture species (i.e., water vapor, cloud water, rain, cloud ice, snow and graupel); the physical processes of rain production, fall and evaporation; the cloud water accretion and autoconversion; condensation; and saturation adjustment and ice sedimentation. The western Mediterranean is affected by air masses of distinct signature (i.e., Saharan, Atlantic, purely Mediterranean or continental central European), featuring a high variability of aerosol concentration that influences the moist physical mechanisms. The inclusion of two different CCN concentrations copes with uncertainties in the aerosol characteristics. On the other hand, the choice of different PBL schemes can be crucial when correctly simulating the onset of mesoscale severe weather phenomena. PBL modulates the temperature and moisture profiles in the lower troposphere and the effects of turbulence in the daytime convective conditions (Hu et al., 2010; Coniglio et al., 2013). Finally, it is worth noting that the initial and lateral boundary conditions are kept invariant through all the MPS ensemble members. IC/LBC come from the ECMWF-EPS reference forecast for each individual case study, and lateral boundary conditions are updated every 3 h.
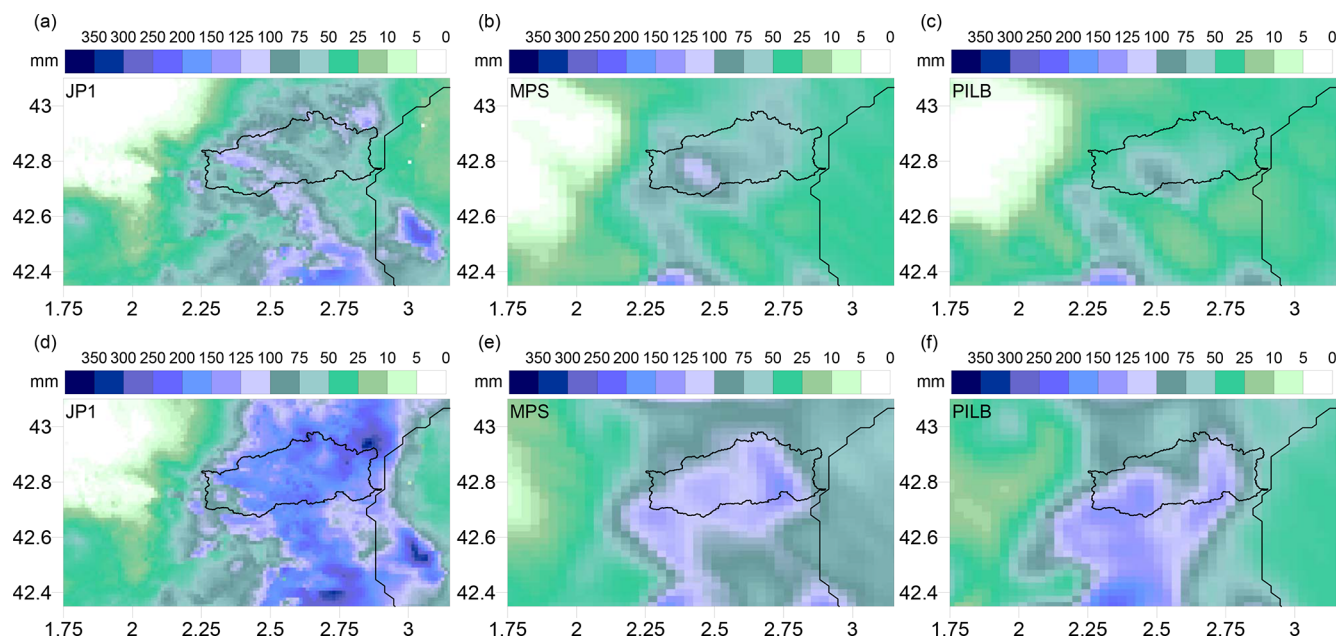
Figure 6. Spatial distributions of the 48 h rainfall amounts for the March 2013 episode according to **(a)** radar JP1, **(b)** MPS percentile 90 and **(c)** PILB percentile 90, starting on 4 March 00:00 UTC; and **(d)** radar JP1, **(e)** MPS percentile 90 and **(f)** PILB percentile 90, starting on 5 March 00:00 UTC. The Agly basin is highlighted.

5　Results and discussion

5.1　Verification of the SREPS

The quantitative comparison of the spatial 48 h accumulated precipitations for the PILB and MPS experiments against the radar estimates provides a quality outlook of the ensemble performance for the selected episodes over the study region. Figures 6–8 indicate realistic spatial distributions for all the study cases: high rainfall accumulations in the upper tail distributions of both ensemble strategies are a good indication of the potential for heavy rainfall. The regional roughed topography (i.e., the pre-Pyrenees, Pyrenees and the Massif Central) is determinant in placing and focusing the probabilistic quantitative precipitation forecasts. Both approaches could succeed in issuing warning alerts before flash-flood scenarios in the region. However, SREPS reliability must be previously checked at basin scales. Flash-flood forecasting over a single medium-sized catchment is a challenging issue as many small-scale atmospheric factors concur in determining the location of deep convection and intense precipitation. A crucial feature in determining correctly the location of the rainfall amounts is to accurately simulate the south to north-easterly low-level moisture maritime flows impinging over the mountainous slopes of the Agly basin.

The 48 h rain-gauge (PLU) and radar-derived (JP1) rainfall amounts have been used to evaluate the forecasting ensemble skill at the relevant hydrological scales. To this end, the cumulative ensemble QPFs have been interpolated to all the available rain gauges and to the pixels of the radar domain shown in Figs. 6 to 8 for each study case (Akima, 1978, 1996; Fig. 9). Most members of the PILB and MPS ensembles exhibit underestimations for the 4–5 March 2013 and 28–29 November 2014 experiments, while they exhibit overestimations for the 16–18 November 2013 simulations. Both strategies do not present remarkable differences in ensemble skill and spread when forecasting the total rainfall amounts (Fig. 10). Root mean squared errors (RMSEs) and correlations (r) are quite similar, indicating a slightly more accurate performance of the MPS or PILB ensemble strategy depending on the case study and the starting day of the experiment.

In addition, the skill of each ensemble strategy in predicting the probability for different accumulations – ranging from light to torrential rainfalls – has been assessed by means of the ROC (receiver operating characteristic) curves. The ROC curve expresses the true hit rate of a probabilistic forecast at different false-alarm rates, while the area under the ROC curve (AUC) quantifies the ability of the ensemble to discriminate between the occurrence or nonoccurrence of an event (Schwartz et al., 2010). ROC curves have been computed by using all the study cases, and the radar-derived (JP1) rainfall accumulations have been employed as the observed baseline. The following 48 h accumulated precipitation thresholds have been considered: 5, 10, 15, 25, 50, 75, 100, 125, 150 and 200 mm. As the forecast probabilities are computed and verified against each pixel within the radar domain shown in Figs. 6 to 8, the statistical sample sums up to 54 145 members (7735 radar grid points times 7 ensemble experiments).
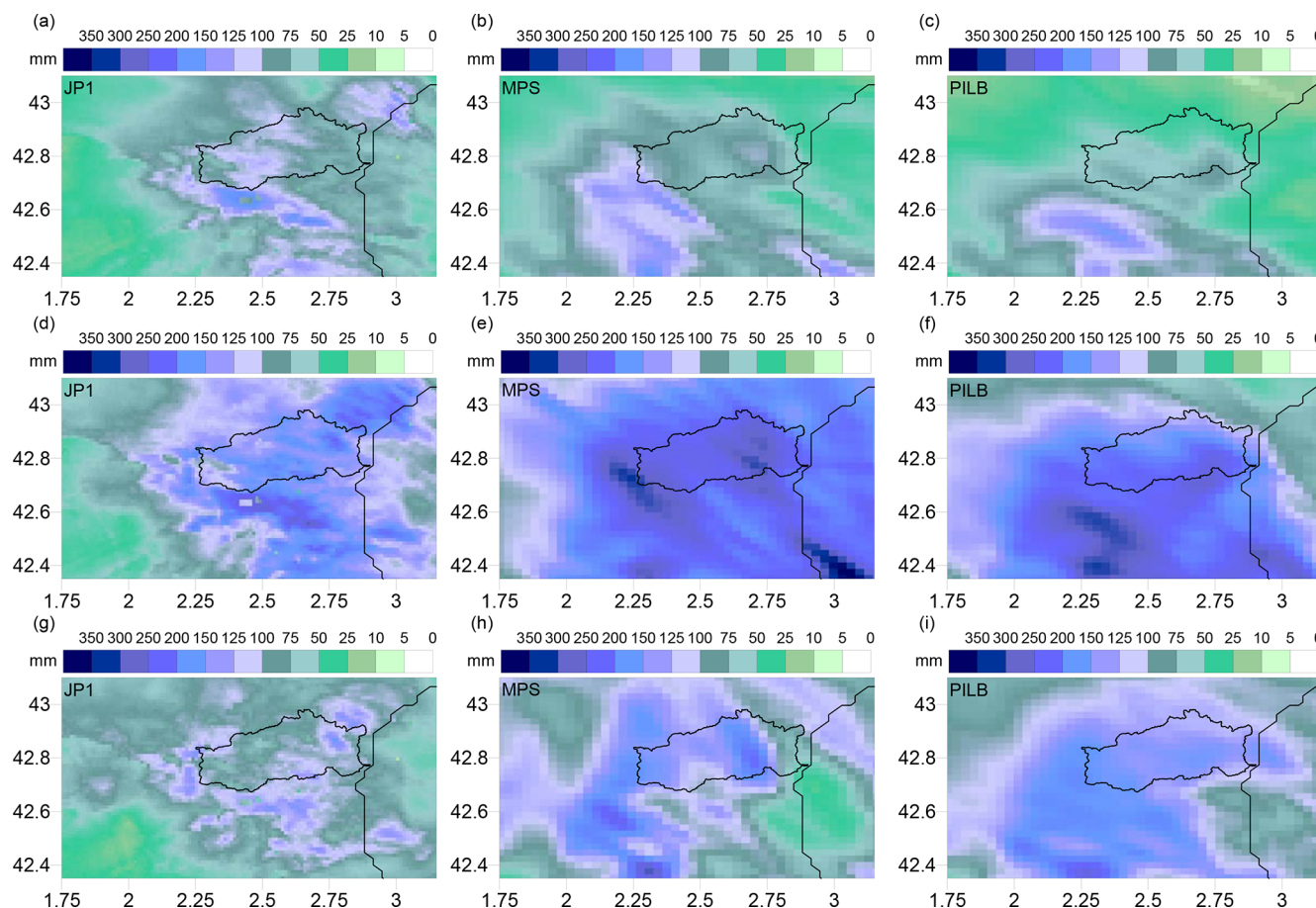
Figure 7. Spatial distributions of the 48 h rainfall amounts for the November 2013 episode according to **(a)** radar JP1, **(b)** MPS percentile 90 and **(c)** PILB percentile 90, starting on 16 November 00:00 UTC; **(d)** radar JP1, **(e)** MPS percentile 90 and **(f)** PILB percentile 90, starting on 17 November 00:00 UTC; and **(g)** radar JP1, **(h)** MPS and **(i)** PILB, starting on 18 November 00:00 UTC. The Agly basin is highlighted.

Probabilistic QPFs from the PILB approach show slightly higher forecasting skills than MPS for small rainfall accumulations (i.e., ≤ 15 mm; Table 5 and Fig. 11). Even so, the AUCs are above 0.85 for both ensemble strategies. For moderate to high rainfall thresholds (25–75 mm), PILB and MPS are almost statistically indistinguishable, with AUCs well above 0.7. Depending on the precipitation limit, MPS or PILB features a slightly higher probabilistic forecasting skill. At greater thresholds (≥ 100 mm), PILB shows a larger discrimination ability, with areas slightly higher than 0.7 for all the cases, except the most extreme precipitation accumulation. On the other hand, MPS renders values close to but below 0.7. In general, both strategies exhibit an elevated quality of the probabilistic forecasts for low to moderate rainfall accumulations. Remarkably, the discrimination ability of the PILB strategy is maintained up to 150 mm. This result points to a more effective encompassing of uncertainties emerging from the IC/LBCs than from the microphysical and PBL physical inaccuracies likely due to the dominant role of the regional complex orography when controlling rainfall location. However, the high AUCs rendered by both ensemble

strategies suggest accounting for both sources of uncertainty so as to obtain high-quality probabilistic quantitative precipitation forecasts.

5.2 Verification of streamflow forecasts

As mentioned by Bellier et al. (2017), the visual inspection of individual hydrographs is useful for a better understanding of how forecasts behave. The hydrological simulations have been forced by the 48 h meteorological simulations, resulting in seven hydrometeorological simulations each lasting 2 d, starting respectively on 4 and 5 March 2013 (20130304_2d and 20130305_2d); 16, 17 and 18 November 2013 (20131116_2d, 2013117_2d and 20131118_2d); and 28 and 29 November 2014 (20141128_2d and 20141129_2d) at 00:00 UTC. Figure 12 shows the hydrographs at three stations (no. 1, no. 2 and no. 5) of the 20130305_2d, 20131117_2d and 20141129_2d experiments and for all the 48 h performed simulations with observed forcing (PLU and JP1), deterministic (WRF) and ensemble forecast MPS. Results are very similar for PILB-HEPS. The
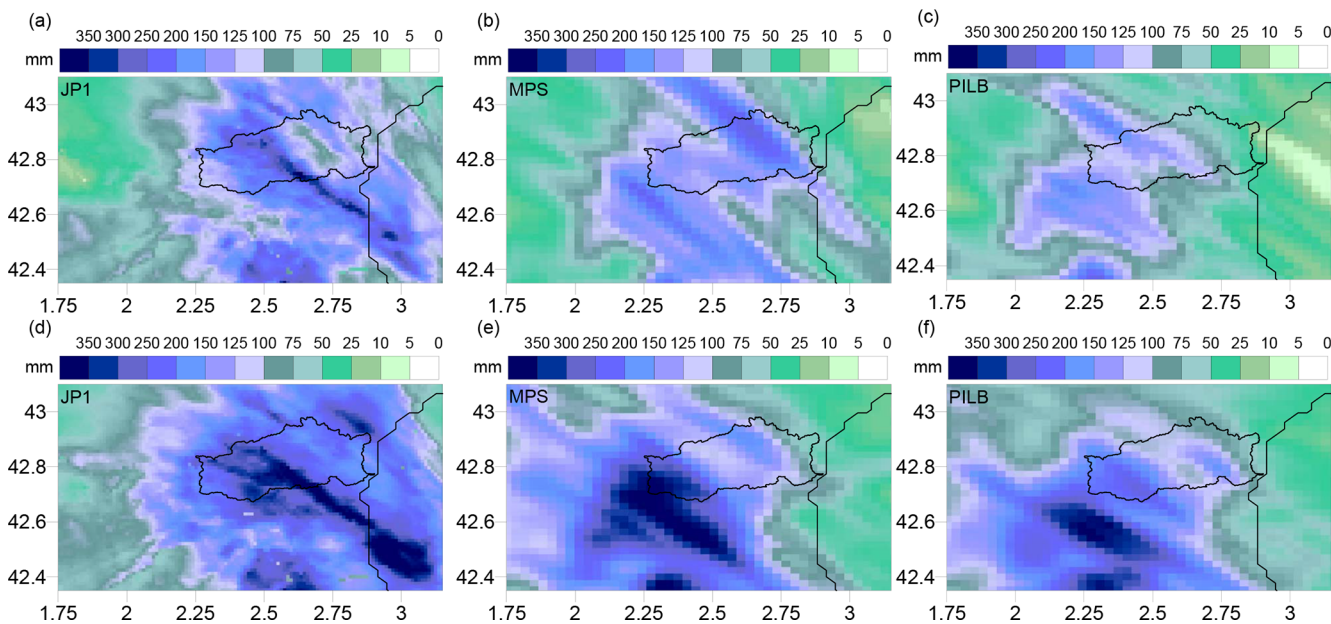
Figure 8. Spatial distributions of the 48 h rainfall amounts for the November 2014 episode according to **(a)** radar JP1, **(b)** MPS percentile 90 and **(c)** PILB percentile 90, starting on 28 November 00:00 UTC; and **(d)** radar JP1, **(e)** MPS percentile 90 and **(f)** PILB percentile 90, starting on 29 November 00:00 UTC. The Agly basin is highlighted.

Table 5. Areas under the ROC curves for the MPS and PILB ensemble strategies. Associated uncertainty of each score (between brackets) is expressed as the 95th percentile confidence intervals, calculated by using a 10 000-sample bootstrap.

Precipitation threshold (mm)	ROC areas	
	MPS	PILB
5	0.855 (0.846–0.864)	0.917 (0.911–0.922)
10	0.888 (0.881–0.894)	0.913 (0.909–0.917)
15	0.852 (0.846–0.859)	0.877 (0.872–0.881)
25	0.833 (0.828–0.839)	0.842 (0.837–0.847)
50	0.785 (0.780–0.790)	0.771 (0.766–0.776)
75	0.741 (0.735–0.746)	0.741 (0.736–0.747)
100	0.699 (0.694–0.705)	0.721 (0.715–0.726)
125	0.690 (0.684–0.695)	0.717 (0.711–0.722)
150	0.691 (0.685–0.697)	0.716 (0.710–0.721)
200	0.638 (0.630–0.647)	0.689 (0.682–0.696)

median and the 10th and 90th quantiles of each ensemble strategy, as well as the first-level alert from the flood warning center in France (SCHAPI), are also shown as references. In general, the WRF deterministically driven hydrological forecasts often miss the peak times for all the hydrometric stations (Fig. 12). The HEPS improves this feature, even if biases in the EPS still remain as they are propagated down to the hydrological model. That is, the MPS-HEPS and PILB-HEPS exhibit slight underestimations (overestimations) for the 20130305_2d and 20141129_2d (20131117_2d) simulations. The observed peak time is included in the boxplots (minimum and maximum of all of the data) of the ensemble strategies for the five stations, whereas it is not included in the boxplot for the deterministic simulations at stations no. 1–3 as it can be seen in Fig. 13 for stations no. 1 and no. 5. It can also be appreciated that the peak timing delay is usually negative, independent of the experimental setup. Almost all the hydrometeorological simulations result in earlier peak timings than observed.

The peak plot approach has been adopted to better appreciate the value of the ensemble strategies: all the ensemble members are joined in a single plot by calculating the deviation from the observed peak discharge and timing (Zappa et al., 2013; Ravazzani et al., 2016). Figures 14–16 summarize the simulations carried out for stations no. 2 and no. 5 and for simulations 20130305_2d, 20131117_2d and 20141129_2d. Results exhibit a high interevent variability as it might be expected given their different characteristics. Regarding the MPS-HEPS experiments, the observed peak lies in the range of variation of the ensemble for the 20130305_2d run at hydrometric stations no. 1 and no. 2 (Fig. 14). This fact can be ascribed to the large spread found in the driven peak discharges: deviations from the observation range from approximately -110 to $+200 \, \mathrm{m^3 \, s^{-1}}$, while timing delays fluctuate from -26 to $+15$ h for station no. 2. Indeed, the 80 % confidence interval of the MPS-HEPS simulations never encompasses the observed discharge for this event. The same remarks also apply for the 20141129_2d case at stations no. 3–5 (Fig. 16) and for 20131117_2d at station no. 3. The 80 % confidence interval of the MPS-HEPS simulations encom-
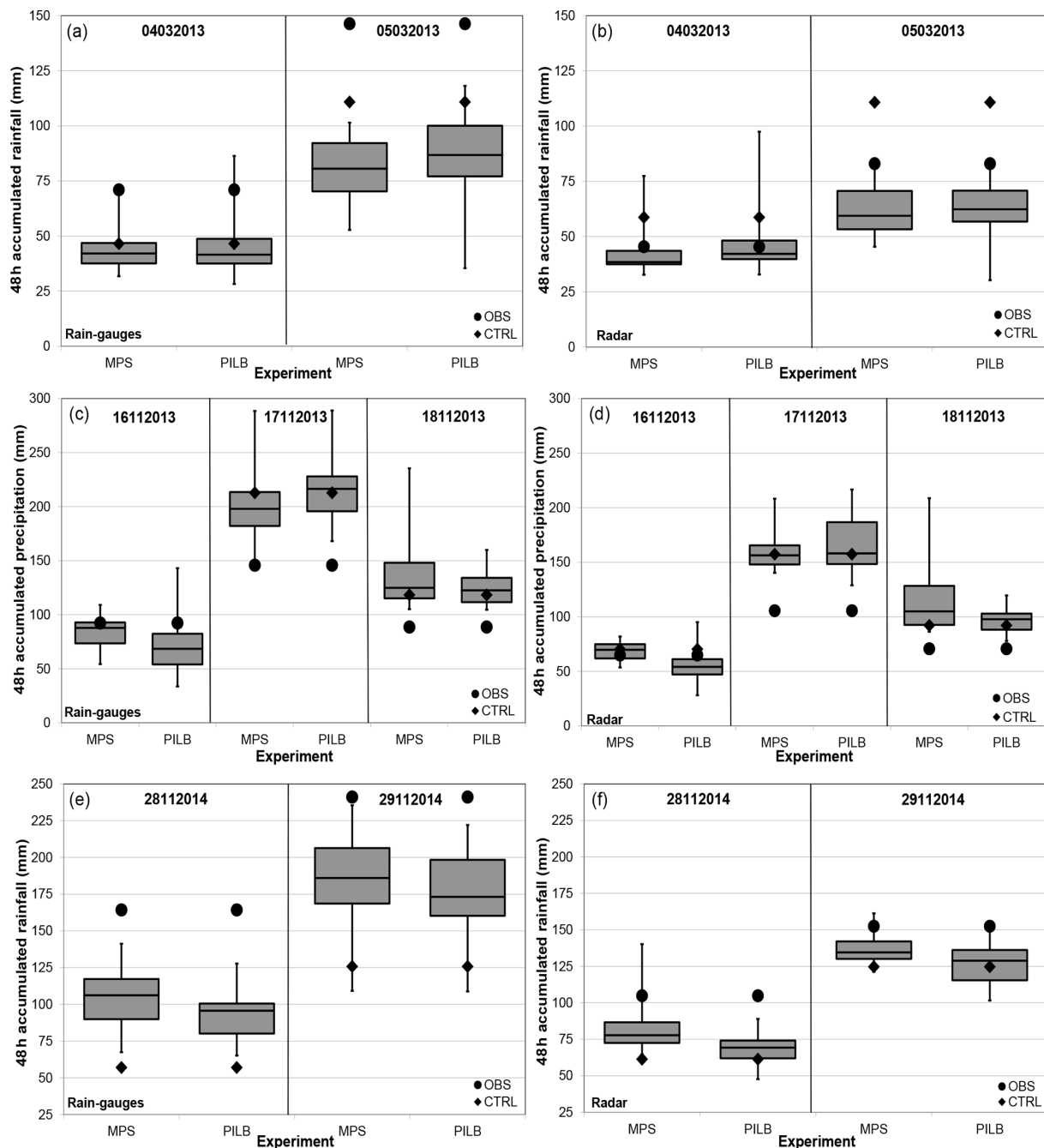
Figure 9. The 48 h rainfall amounts according to the rain-gauge (PLU, **a, c, e**) and radar-derived (JP1, **b, d, f**) observations and the PILB and MPS experiments. Boxes denote the p_{25} and p_{75} interquartile ranges, middle horizontal lines show the ensemble median, and whiskers display the tails of the ensemble. Note that the PILB and MPS ensemble experiments start on the day indicated in the upper part of each subpanel. CTRL stands for the control or deterministic simulation.

passes the observed discharge only for the 20131117_2d simulation at stations no. 2, 4 and 5 (Fig. 15) and for the 20141128_2d simulation at station no. 2.

The observed peak also lies in the range of variation of the PILB-HEPS ensemble strategy for the 20131117_2d run at stations no. 2–5 (Fig. 15) and for the 20141129_2d sim-

ulation at the five gauge stations (Fig. 16). Concerning both episodes at gauge station no. 2, the PILB-HEPS spread is larger than MPS-HEPS in terms of the observed peak discharge although smaller for the observed peak time. That is, from −17 to +22 h for the MPS-HEPS and from −3 to +18 h for the PILB-HEPS for 20131117_2d as well as from
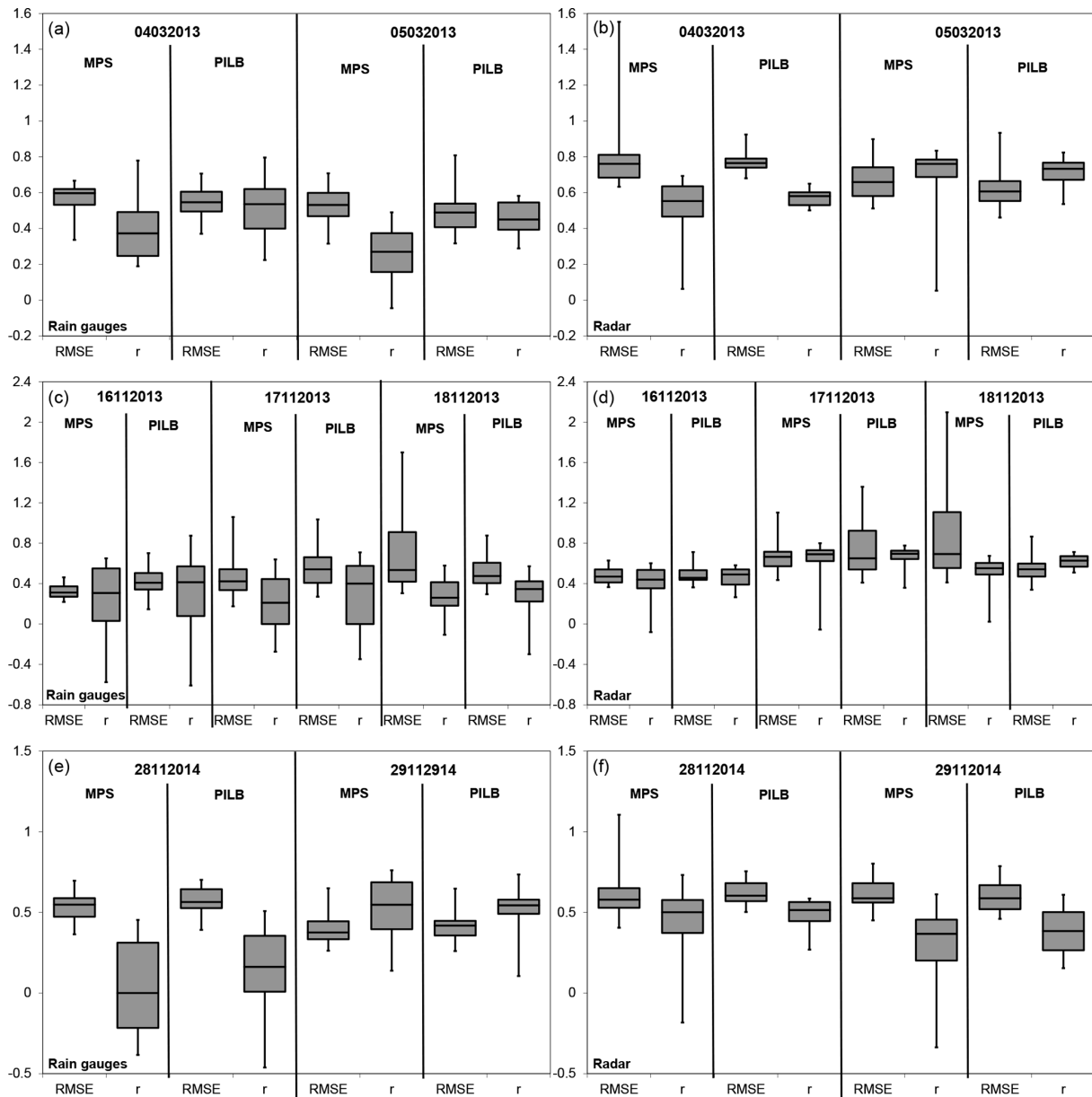
Figure 10. Statistical scores of the 48 h rainfall amounts for the PILB and MPS ensemble members when compared against the rain-gauge (PLU, **a, c, e**) and the radar-driven (JP1, **b, d, f**) observations. Boxes denote the p_{25} and p_{75} interquartile ranges, middle horizontal lines show the ensemble median, and whiskers display the best and the worst ensemble members. Note that the PILB and MPS ensembles start on the day indicated in the upper part of each subpanel.

-12 to $+25$ h for the MPS-HEPS and from -12 to $+8$ h for the PILB-HEPS for 20141129_2d. The opposite is found at station no. 5 for 20130305_2d and 20141129_2d. The 80 % confidence interval of the PILB-HEPS simulations encompasses the observed discharge only for the 20141128_2d run at station no. 2 and for the 20141129_2d run at stations no. 2 and 3 (Fig. 16). Given those results, it seems that there are no substantial differences between the both HEPS strategies on these test cases.

5.3 System reliability for flood warning

Results of all the performed hydrometeorological simulations lead to the conclusion that it is very difficult to correctly reproduce the spatial variability of the catchment behavior, even forcing the hydrological model with observed rainfall. The next step was therefore to test the ability of the hydrometeorological modeling strategies in issuing reliable flood warnings.
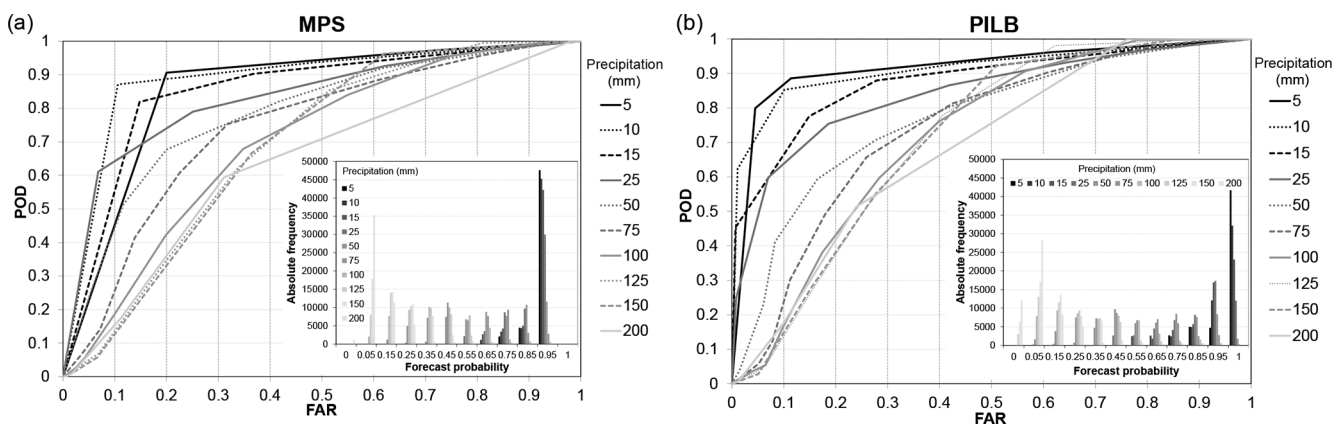
Figure 11. ROC curves of the MPS and PILB ensemble strategies. The embedded figures display the sharpness diagrams containing the number of forecasts used in each probability bin and the total number of observations considered.

Table 6. Two-by-two contingency table for flood warning evaluation.

		Threshold exceedance observed	
		Yes	No
Threshold exceedance forecast	Yes	Hits (h)	False alarms (f)
	No	Misses (m)	Correct negatives (n)

Let us consider a forecast event that either occurs or does not occur. For flood forecasting, it usually consists in an alert threshold exceedance. The performance of a hydrometeorological prediction chain can be examined using a contingency table (Table 6).

Several metrics for the evaluation of flood warning performance can be derived from the contingency table by considering the number of hits (h), misses (m), false alarms (f) and correct negatives (n) for all the simulations. The proportion correct (PC), probability of detection (POD), false-alarm ratio (FAR), critical success index (CSI) and BIAS have the following properties (Nurmi, 2003):

– The PC score corresponds to the ratio of correct warning forecasts and total forecasts. PC ranges from 0 to 1, with the latter being the perfect score. Note that the PC index does not differentiate between misses and false alarms.

– The probability of detection is the ratio of correctly forecast threshold exceedances to the total number of threshold exceedance observed. POD ranges from 0 (no hit) to 1, with 1 being the best. Note that for values equal to 1, there are no misses and all occurrences of the event were correctly forecast. However, POD does not penalize false alarms, and it can be artificially improved by overforecasting.

– The false-alarm ratio is the ratio of the number of false alarms to the total number of threshold exceedance forecasts. FAR ranges from 0 to 1, with 0 being perfect. That

is, there are no false alarms and all warning forecasts were correct. Note that FAR does not penalize misses, and it can be artificially improved by underforecasting.

– Neither POD nor FAR can give a complete picture of forecasting success. The critical success index combines both aspects of the probability of detection and false-alarm ratio. Therefore, CSI is more balanced and better quantifies the correspondence between the observed and forecasted occurrences. This index is sensitive to hits and penalizes both misses and false alarms. CSI values range from 0 (no hit) to 1 (no misses, no false alarms), with 1 being the best. CSI ignores correct negatives as what is expected in the forecast is to be effective in case of alert.

– The frequency bias compares the number of times an event was forecast to the number of times an event was observed. If BIAS = 1, both frequencies are equal and the forecast is unbiased. If BIAS > 1 (< 1), there is an overforecast (underforecast) tendency: the event was forecast more (less) than it was observed.

As a first step, the probability of exceeding the warning threshold has been calculated for each ensemble strategy. The warning threshold that is used here is the first-level alert from the flood warning center in France (SCHAPI) as plotted in Fig. 12. Results are very similar for MPS-HEPS and PILB-HEPS: overall, with respect to the deterministic simulations, both ensemble strategies improve the forecast of threshold
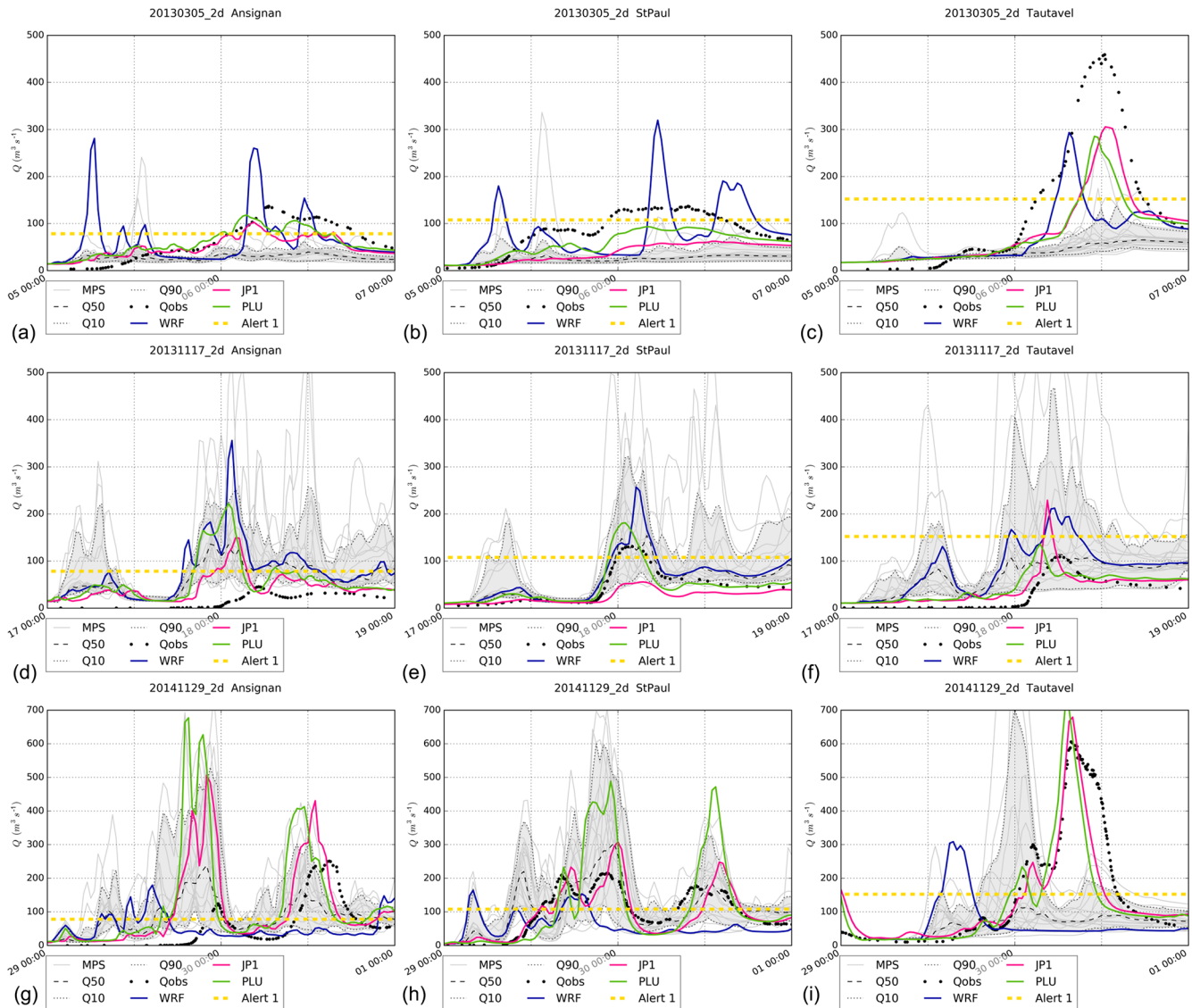
Figure 12. MPS-HEPS hydrographs at station no. 1 **(a, d, g)**, no. 2 **(b, e, h)** and no. 5 **(c, f, i)** for the 20130305_2d simulation **(a–c)**, 20131117_2d simulation **(d–f)** and 20141129_2d simulation **(g–i)**. Note that Q_{50} is the ensemble median, Q_{10} denotes the 10th ensemble quantile, Q_{90} labels the 90th ensemble quantile, Q_{obs} is the observed discharge, WRF is the WRF deterministically driven discharge experiment, PLU is the PLU-driven runoff simulation and JP1 denotes the JP1-driven discharge simulation. Alert 1 corresponds to the first-level alert.

exceedance for station no. 5 (Tautavel) and degrade it for station no. 2 (Saint-Paul-de-Fenouillet), whereas there is no clear trend for station no. 1 (Ansignan). As it has been stated in Sect. 3.2, when the hydrologic simulations are suitable for the eastern Agly (station no. 2), the discharge is overestimated over the western part (station no. 5). As most members of the PILB and MPS ensembles exhibit underestimations for the 4–5 March 2013 and 28–29 November 2014 events, both MPS-HEPS and PILB-HEPS result in less false alarms for station no. 5 and more misses for station no. 2. PILB and MPS ensembles also exhibit overestimations for the 16–18 November 2013 event but less than the determin-

istic simulation; results are therefore the same as for the two other events.

Figures 17 to 19 show the results for FAR, CSI and BIAS scores at the five hydrometric sections. These scores are calculated with respect to the observed discharges and by using all the runs of the different episodes. As 48 h simulations have been performed, these scores are based on the following seven experiments described in Sect. 5.2: 20130304_2d, 20130305_2d, 20131116_2d, 2013117_2d, 20131118_2d, 20141128_2d and 20141129_2d. Some tendencies can be highlighted from these results.
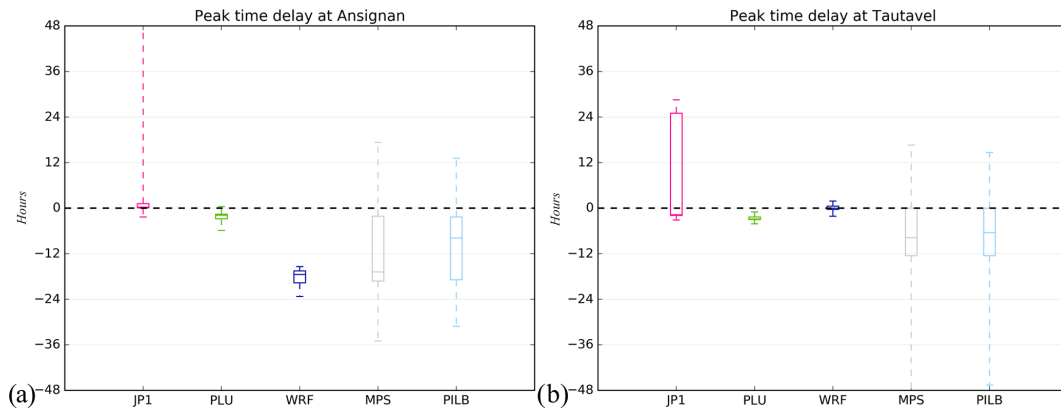
Figure 13. Delay of simulated peak time for the seven simulations at stations no. 1 **(a)** and no. 5 **(b)** for simulations with JP1 forcing, PLU forcing, WRF deterministic forcing and ensemble strategies' forcings (MPS and PILB). The boxplot presents five sample statistics: the minimum, the lower quartile, the median, the upper quartile and the maximum.
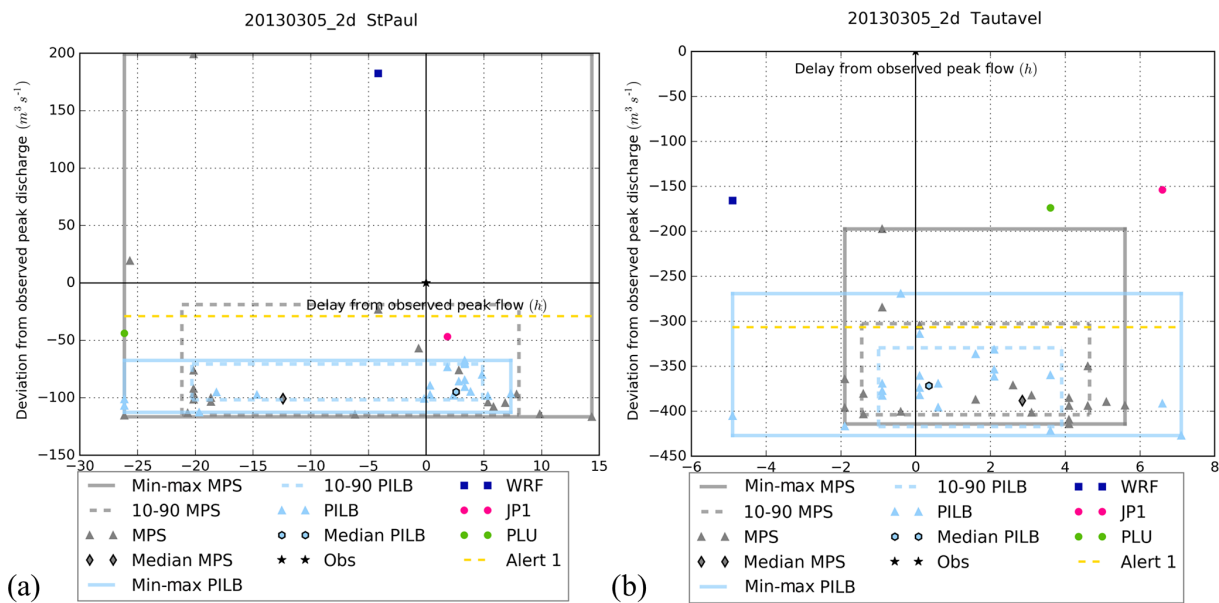

Figure 14. Peak flow analysis at stations no. 2 **(a)** and no. 5 **(b)** for 20130305_2d. The *x* axis shows the delay from the observed peak time; the *y* axis shows the deviation from the observed peak discharge. The triangles show the deviation of the simulations with ensemble member forcing (grey for MPS, light blue for PILB), the shapes with black contour show the deviation of the median of the HEPS simulations with ensemble member forcing, the pink circle shows the deviation of the simulation with JP1 forcing, the green circle shows the deviation of the simulation with PLU forcing and the dark blue square shows the deviation of the simulation with deterministic WRF forcing. Alert 1 (yellow dashed line) is the warning threshold; the black star is the observation used as normalized reference.

- The MPS-HEPS strategy overall performs better than the PILB-HEPS approach for the tested scores. However, both ensemble strategies' scores are very similar.

- No ensemble strategy performs best for station no. 2 for FAR and CSI: there is no false alarm at this station (Fig. 17), and therefore the CSI score is the best with respect to the other stations (Fig. 18).

- Although the ensemble improves the peak timing in some events, it does not improve the issuance of warn-

ing, at least according to the five tested scores: the deterministic WRF simulation always has better scores than the median of both MPS-HEPS and PILB-HEPS, except for BIAS, and sometimes better than the maximum.

BIAS shows that both ensemble strategies tend to underestimate the discharge at all the gauge stations except station no. 1, in the mountainous part of the catchment (Fig. 19). That is, MPS-HEPS and PILB-HEPS tend to underestimate the discharge at all the stations except over the mountainous part of the catchment. This is an indication of the paramount
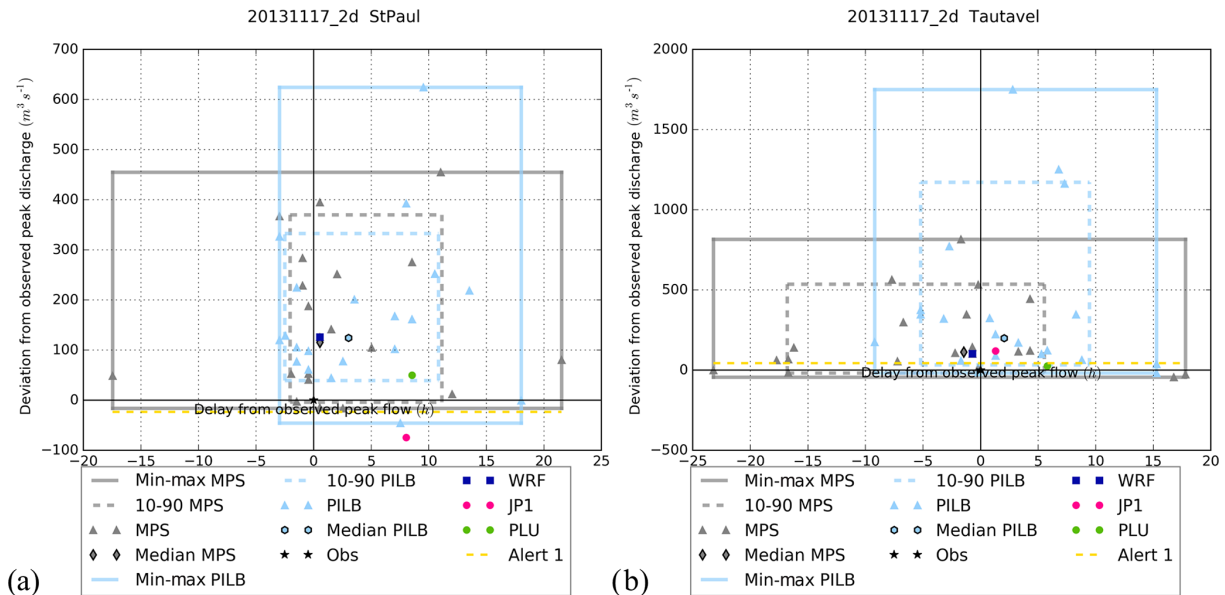
Figure 15. Peak flow analysis at stations no. 2 **(a)** and no. 5 **(b)** for 20131117_2d. See Fig. 14 for the details of the legend.
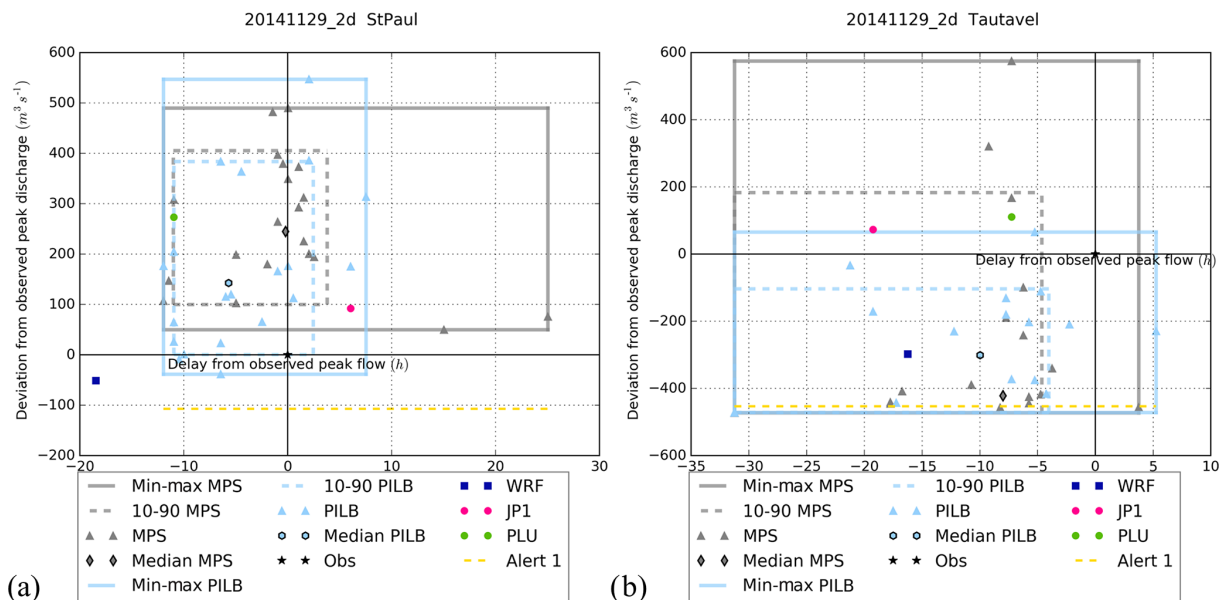

Figure 16. Peak flow analysis at stations no. 2 **(a)** and no. 5 **(b)** for 20141129_2d. See Fig. 14 for the details of the legend.

importance of the orography when controlling the location of deep convection in the meteorological simulations. When orography does not play such an important role, forecasting the small-scale atmospheric features linked to the triggering and development of highly localized convective precipitation cores is more uncertain. As mentioned before, PILB-HEPS and MPS-HEPS tend to exhibit underestimations for both 20130305_2d and 20141129_2d simulations and overestimations for the 20131117_2d run. Conversely, the observed forcing and the deterministic forecast tend to overestimate the discharge, except for the two eastern stations, no. 4 and

no. 5. We find here the consequences of the hydrological model calibration: when the simulated hydrographs are suitable for the eastern Agly, the discharge is overestimated over the western part (Sect. 3.2).

Quantitative discharge forecasts can be evaluated against observed discharges but also against simulated discharges using observed forcings. As stated by several authors (Verkade et al., 2013; Bellier et al., 2017), the errors due to the parameters and structure of the hydrologic model are therefore not taken into account in the last case. This approach separates the impact of the external-scale uncertainties from

Figure 17. False-alarm ratio (FAR) scores at the five gauging stations for the seven simulations. Statistical indices have been computed by using the observed discharge. Experiments are labeled as follows. WRF: simulated discharge with deterministic WRF forcing; PLU: simulated discharge with PLU forcing; JP1: simulated discharge with JP1 forcing; MPS and PILB: ensemble strategies. The boxplot presents five sample statistics: the minimum, the lower quartile, the median, the upper quartile and the maximum.

Figure 18. As Fig. 17 but for CSI.

Figure 19. As Fig. 17 but for BIAS.

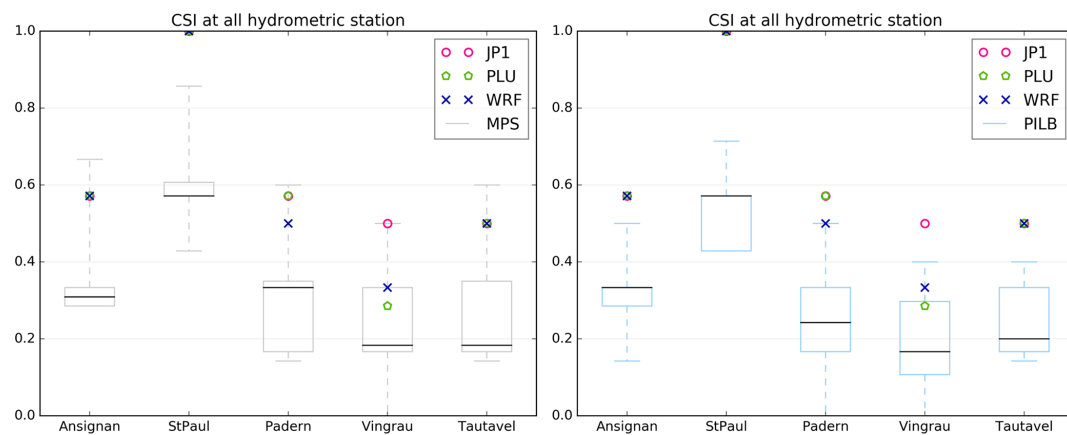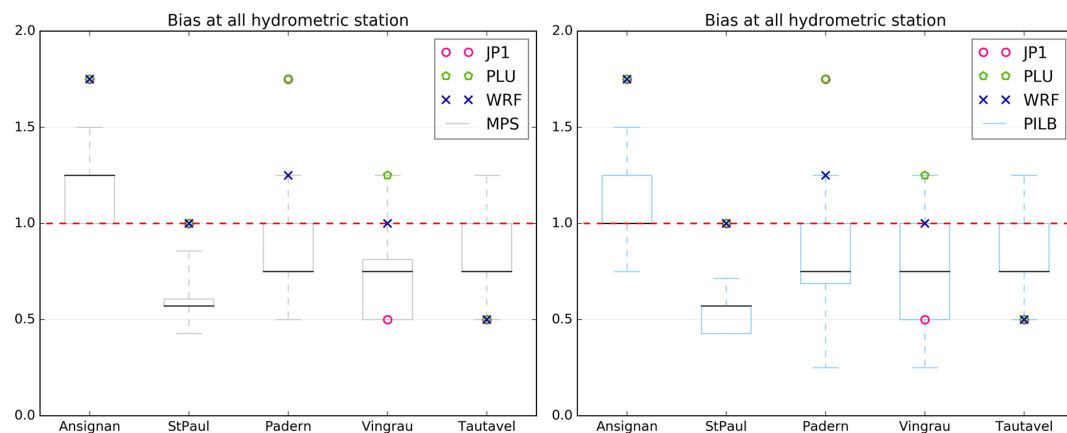those emerging from the hydrological model. Evaluations have been again performed by using the simulated discharges with observed forcing PLU and JP1 as the baseline instead of the observed flows.

As expected, when only external-scale uncertainties are taken into account, the scores for the evaluation against simulated discharges with PLU or JP1 improve: PC, POD and CSI are higher, and there are no false alarms at three stations (no. 1–no. 3). However, the BIAS score shows that both ensemble strategies tend to highly underestimate the simulated discharge at all the stations, except at station no. 5 when compared to PLU and at stations no. 4 and no. 5 when compared to JP1 (Fig. 20). These stream gauges are located over the eastern part of the catchment. Again, the deterministic WRF simulations have better scores than the median of both HEPS, except for station no. 4, and the PC, POD, FAR and BIAS scores when compared to JP1.

5.4　Overall view of the modeling performance

Binary events highlight one aspect of the forecast, which is especially relevant to avoid casualties, damages or economic losses (Hersbach, 2000). To obtain a more general quantification of the ensemble performances, other criteria are necessary. Here, the overall discharge forecast at the five gaging stations is studied by using the continuous rank probability score (CRPS; Matheson and Winkler, 1976). The CRPS measures the differences between the forecast, $P(x)$, and observation, $P_a(x)$, expressed as cumulative distributions of one parameter x (Eq. 4). This score has the dimension of the parameter and is equal to the mean absolute error (MAE) for a deterministic forecast. The following description is mainly retrieved from Hersbach (2000):

$$\text{CRPS} = \int_{-\infty}^{+\infty} [P(x) - P_a(x)]^2 dx, \tag{4}$$

where x is the parameter of interest, herein the discharge, and x_a is the value that actually occurred. $P(x)$ and $P_a(x)$ are the cumulative distributions of x and x_a, respectively (Eqs. 5 and 6).

$$P(x) = \int_{-\infty}^{x} \rho(y) dy, \tag{5}$$

where $\rho(x)$ is the probability density function of the forecast x.

$$P_a(x) = H(x - x_a) = \begin{cases} 0 & \text{for} \quad x < x_a, \\ 1 & \text{for} \quad x \geq x_a, \end{cases} \tag{6}$$

where H is the Heaviside function. The minimum value of the CRPS is zero for a perfect deterministic forecast (i.e., $P(x) = P_a(x)$).

Herein, the CRPS is averaged over the ensemble members and is therefore noted $\overline{\text{CRPS}}$, while the x parameter corresponds to the discharge at the five gaging stations. The $\overline{\text{CRPS}}$ is very small for the simulations corresponding to the episode of November 2013 (i.e., 20131116_2d, 20131117_2d and 20131118_2d). This score is always below $10\,\text{m}^3\,\text{s}^{-1}$ for all stations and the MPS-HEPS and PILB-HEPS strategies. Conversely, the $\overline{\text{CRPS}}$ is quite high – above $50\,\text{m}^3\,\text{s}^{-1}$ – for the numerical runs of the event of November 2014 (i.e., 20141128_2d and 20141129_2d), especially at station no. 5. That is, the cumulative distributions of discharge are similar between the HEPSs and the observed discharges for the event of November 2013, but they are dissimilar for the episode of November 2014. Concerning the experiments for the episode of March 2013 (i.e., 20130304_2d and 20130305_2d), the $\overline{\text{CRPS}}$ is low for stations no. 1 and no. 3 (below $15\,\text{m}^3\,\text{s}^{-1}$) and higher for stations no. 2, no. 4 and no. 5 (close to or above $20\,\text{m}^3\,\text{s}^{-1}$).

To evaluate more easily the performances of the ensemble strategies, their performances are also compared against the efficiency of a reference forecast by using the skill score with respect to the $\overline{\text{CRPS}}$ (Eq. 7) (Bontron, 2004):

$$\text{CRPSS} = 1 - \frac{\overline{\text{CRPS}}}{\text{CRPS}_{\text{ref}}}. \tag{7}$$

The chosen reference forecast is the simulation performed with the deterministic forecast (WRF), and in that case the $\overline{\text{CRPS}}$ skill score is written as follows:

$$\text{CRPSS} = 1 - \frac{\overline{\text{CRPS}}}{\text{MAE(WRF)}}. \tag{8}$$

A CRPSS of 1 corresponds to a perfect forecast ($\overline{\text{CRPS}} = 0$), while a value of 0 indicates that the HEPS and the reference forecast have the same performances ($\overline{\text{CRPS}} = \text{MAE(WRF)}$). Negative skill scores denote that the reference prediction performs better than the HEPS ($\overline{\text{CRPS}} > \text{MAE(WRF)}$).

Figure 21 shows that the two ensemble strategies exhibit very similar skill score CRPSS:

- in general, both ensemble strategies perform better than the deterministic WRF experiment, except for 20130304_2d and 20130305_2d;

- the main differences between both ensemble strategies are found for the 20131118_2d experiment – PILB clearly outperforms MPS at all the stream stations.

As stated before, selecting the runoff simulation driven by the deterministic weather forecast as the reference does not account for the errors due to the hydrological model. The $\overline{\text{CRPS}}$ skill score can also be calculated by using the simulation performed with the observed precipitation fields (PLU and JP1) as the reference:
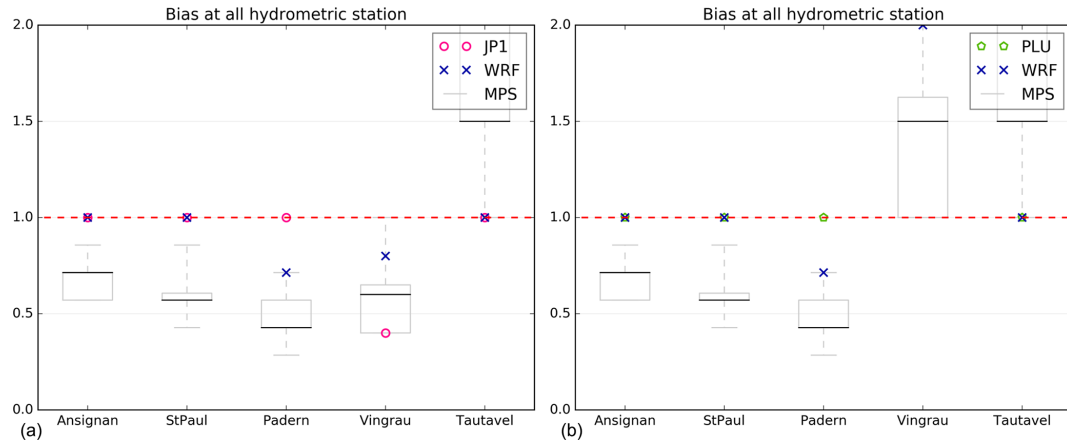
Figure 20. Bias scores with respect to the simulated discharges with forcing PLU **(a)** and forcing JP1 **(b)** at the five gaging stations for all the simulations of the seven simulations. WRF: simulated discharge with deterministic WRF forcing; PLU: simulated discharge with PLU forcing; JP1: simulated discharge with JP1 forcing; MPS: ensemble strategies. The boxplot presents five sample statistics: the minimum, the lower quartile, the median, the upper quartile and the maximum.
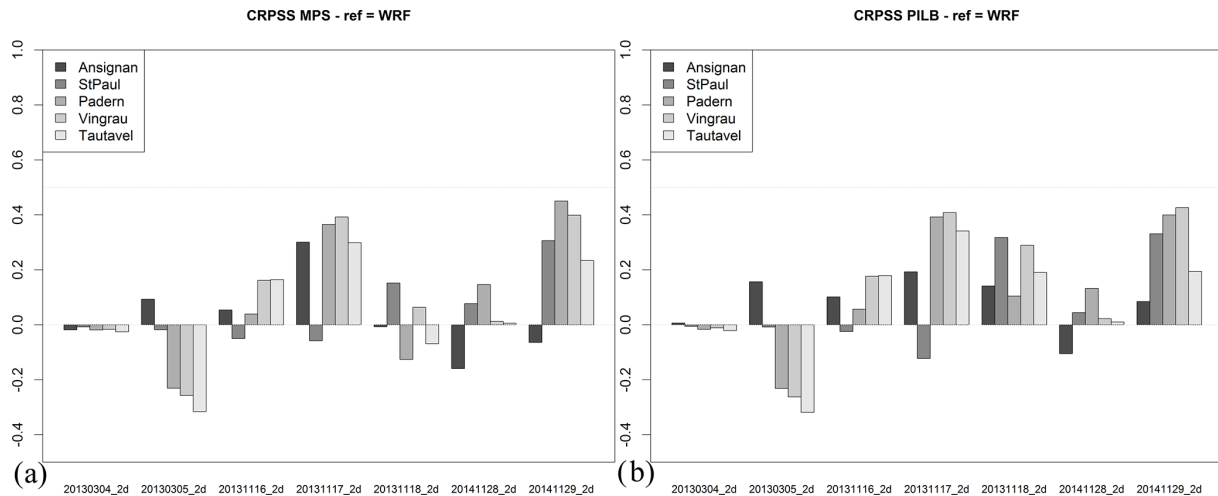

Figure 21. $\overline{\text{CRPS}}$ skill scores of the seven 48 h experiments and at the five hydrometric stations for the **(a)** MPS-HEPS and **(b)** PILB-HEPS strategies. The reference forecast is the deterministic WRF experiment.

$$\text{CRPSS}_{\text{PLU}} = 1 - \frac{\overline{\text{CRPS}}}{\text{MAE(PLU)}},$$

$$\text{CRPSS}_{\text{JP1}} = 1 - \frac{\overline{\text{CRPS}}}{\text{MAE(JP1)}}. \qquad (9)$$

Not surprisingly, both ensemble strategies have an overall lower performance when compared with the PLU- and JP1-driven runoff simulations, except for event of November 2013. It is interesting to note that for the 20131118_2d run, the PILB-driven runoff forecasts outperform the radar-driven discharge simulation (Fig. 22, right). This is consistent with the previous analyses: events with relatively moderate peak discharge – as the event of November 2013 – are not correctly simulated by MARINE regardless of the observed forcing (Table 4), whereas the $\overline{\text{CRPS}}$ is very low for

the ensemble simulations of the event of November 2013. As stated before, a low $\overline{\text{CRPS}}$ means that the cumulative distributions of discharge are similar between both HEPSs and the observed discharges for the event of November 2013, but they are dissimilar between the simulations with both observed forcings and observed discharges for the same event. This may be related to the fact that MPS-HEPS and PILB-HEPS exhibit overestimations for this event, maybe compensating for errors in the model structure that prevent the simulation with observed forcings for this event to be efficient. Both ensemble strategies outperform the hydrological simulations driven by observed forcings (PLU and JP1) for the mountainous station (no. 1: Ansignan) and the 20141128_2d, 20141129_2d, 20131116_2d and 20131118_2d runs. This result is consistent with the difficulty in obtaining satisfac-

Figure 22. As Fig. 21 but just for the PILB-HEPS and the **(a)** PLU and **(b)** JP1 as reference.

tory observations of rainfall in mountainous areas owing to sparse rain-gauge deployment and beam radar blockage.

6 Conclusion

One of the main scientific aims of the HyMeX program is to improve the hydrometeorological forecasting of flash floods over the western Mediterranean region. To this end, three of the most important floods that recently developed over the Agly basin have been selected as study cases. Flood forecasting is a challenging task over this region: high spatial and temporal variability in convective cores and rainfall intensity, strong nonlinearities in the rainfall-runoff transformation, and antecedent moisture conditions lead to a myriad of hydrological responses. This work has focused on coping with uncertainties emerging from the initial and lateral boundary conditions and formulation of numerical weather prediction models. To this end, potentialities of MPS-HEPS and PILB-HEPS ensembles have been examined so as to produce suitable flood forecasts over the Agly basin. The main conclusions are as follows.

– A better ensemble generation strategy at the regional scale has not been found. Similarities in the performance of the MPS and PILB approaches indicate that both sources of external-scale uncertainty contribute similarly to produce adequate levels of skill and spread in the probabilistic quantitative precipitation forecasts.

– Ensemble hydrometeorological simulations have turned out to be satisfactory for alarm detection, even if individual ensemble members can be far from the observations. Alarm systems benefit from large hydrometeorological ensemble spreads.

– The overall HEPS performances improved the deterministically driven runoff simulations.

Some unexpected results also raise interesting questions. For instance, the November 2013 event was poorly simulated using both observed forcings, but ensemble strategies improved the overall discharge forecast. What is the specificity of the November 2013 event that makes it poorly simulated? Is it due to the radar and rain-gauge location or to the initial state of the catchment? Is it due to the model structure itself, which does not represent all the hydrological processes involved (karstic system and snowmelt mainly)? These issues require further investigations and probably more test cases. The next logical approach will be to estimate the uncertainties in the hydrological modeling. Performing hydrological model ensembles to test the errors due to the model calibration is time-consuming. However, according to Douinot et al. (2017), it is also useful in identifying the strengths and weaknesses of the model when simulating the distinct hydrological processes. Hopefully, the future implementation of a hydrological model ensemble will provide the beginning of the answers to the above questions.

Author contributions. HR, AA and RR provided the theoretical background, designed the methodology and analyzed the data. AA ran the meteorological model and wrote the corresponding part. HR ran the hydrological model and wrote the corresponding part.

References

Agence de l'eau Rhône Méditerranée & Corse: Étude de détermination des volumes prélevables, Bassin versant de l'Agly, Technical report, available at: http://www.pyrenees-orientales.gouv.fr/content/download/9251/55322/file/FINAL+Phase+1A3.pdf (last access: 20 November 2019), 2012.

Akima, H.: A method of bivariate interpolation and smooth surface fitting for irregularly distributed data points, ACM Trans. Math. Softw., 4, 148–164, https://doi.org/10.1145/355780.355786, 1978.

Akima, H.: Algorithm 761: Scattered-data surface fitting that has the accuracy of a cubic polynomial, ACM Trans. Math. Softw., 22, 362–371, https://doi.org/10.1145/232826.232856, 1996.

Amengual, A., Romero, R., and Alonso, S.: Hydro-meteorological ensemble simulations of flood events over a small basin of Majorca Island, Spain, Q. J. Roy. Meteorol. Soc., 134, 1221–1242, 2008.

Amengual, A., Carrió, D. S., Ravazzani, G., and Homar, V.: A Comparison of Ensemble Strategies for Flash Flood Forecasting: The 12 October 2007 Case Study in Valencia, Spain, J. Hydrometeo-

rol., 18, 1143–1166, https://doi.org/10.1175/JHM-D-16-0281.1, 2017.

Angevine, W. M., Jiang, H., and Mauritsen, T.: Performance of an eddy diffusivity–mass flux scheme for shallow cumulus boundary layers, Mon. Weather Rev., 13, 2895–2912, https://doi.org/10.1175/2010MWR3142.1, 2010.

Antonetti, M., Horat, C., Sideris, I. V., and Zappa, M.: Ensemble flood forecasting considering dominant runoff processes – Part 1: Set-up and application to nested basins (Emme, Switzerland), Nat. Hazards Earth Syst. Sci., 19, 19–40, https://doi.org/10.5194/nhess-19-19-2019, 2019.

Bartholmes, J. C., Thielen, J., Ramos, M. H., and Gentilini, S.: The european flood alert system EFAS – Part 2: Statistical skill assessment of probabilistic and deterministic operational forecasts, Hydrol. Earth Syst. Sci., 13, 141–153, https://doi.org/10.5194/hess-13-141-2009, 2009.

Bellier, J., Bontron, G., and Zin, I.: Using Meteorological Analogues for Reordering Postprocessed Precipitation Ensembles in Hydrological Forecasting, Water Resour. Res., 53, 10085–10107, https://doi.org/10.1002/2017WR021245, 2017.

Bellier, J., Zin, I., and Bontron, G.: Generating Coherent Ensemble Forecasts After Hydrological Postprocessing: Adaptations of ECC-Based Methods, Water Resour. Res., 54, 5741–5762, https://doi.org/10.1029/2018WR022601, 2018.

Benninga, H.-J. F., Booij, M. J., Romanowicz, R. J., and Rientjes, T. H. M.: Performance of ensemble streamflow forecasts under varied hydro-meteorological conditions, Hydrol. Earth Syst. Sci., 21, 5273–5291, https://doi.org/10.5194/hess-21-5273-2017, 2017.

Beven, K. J. and Kirkby, M. J.: A physically based, variable contributing area model of basin hydrology, Hydrolog. Sci. J., 24, 43–69, https://doi.org/10.1080/02626667909491834, 1979.

Bontron, G.: Prévision quantitative des précipitations: adaptation probabiliste par recherche d'analogues. Utilisation des réanalyses NCEP/NCAR et application aux précipitations du Sud-Est de la France, PhD Thesis from Institut National Polytechnique de Grenoble, France, available at: https://tel.archives-ouvertes.fr/tel-01090969 (last access: 31 January 2020), 2004.

Braud, I., Ayral, P.-A., Bouvier, C., Branger, F., Delrieu, G., Le Coz, J., Nord, G., Vandervaere, J.-P., Anquetin, S., Adamovic, M., Andrieu, J., Batiot, C., Boudevillain, B., Brunet, P., Carreau, J., Confoland, A., Didon-Lescot, J.-F., Domergue, J.-M., Douvinet, J., Dramais, G., Freydier, R., Gérard, S., Huza, J., Leblois, E., Le Bourgeois, O., Le Boursicaud, R., Marchand, P., Martin, P., Nottale, L., Patris, N., Renard, B., Seidel, J.-L., Taupin, J.-D., Vannier, O., Vincendon, B., and Wijbrans, A.: Multiscale hydro-meteorological observation and modelling for flash flood understanding, Hydrol. Earth Syst. Sci., 18, 3733–3761, https://doi.org/10.5194/hess-18-3733-2014, 2014.

Buizza, R. and Palmer, T. N.: The singular-vector structure of the atmospheric general circulation, J. Atmos. Sci. Eng., 52, 1434–1456, https://doi.org/10.1175/1520-0469(1995)052<1434:TSVSOT>2.0.CO;2, 1995.

Champeaux, J.-L., Dupuy, P., Laurantin, O., Soulan, I., Tabary, P., and Soubeyroux, J.-M.: Rainfall measurements and quantitative precipitation estimations at Météo-France: inventory and prospects, La Houille Blanche, 5, 28–34, https://doi.org/10.1051/lhb/2009052, 2009.

Cloke, H. L. and Pappenberger, F.: Ensemble flood forecasting: A review, J. Hydrol., 375, 613–626, https://doi.org/10.1016/j.jhydrol.2009.06.005, 2009.

Cloke, H. L., Pappenberger, F., van Andel, S. J., Schaake, J., Thielen, J., and Ramos, M.-H. (Eds.): Special Issue on Hydrological Ensemble Prediction Systems (HEPS), Hydrol. Process., 27, 1–163, 2013.

Coniglio, M. C., Correia Jr., J., Marsh, P. T., and Kong, F.: Verification of convection-allowing WRF Model forecasts of the planetary boundary layer using sounding observations, Weather Forecast., 28, 842–862, https://doi.org/10.1175/WAF-D-12-00103.1, 2013.

DIREN Languedoc-Roussillon/SIEE-GINGER: Atlas des zones inondables du bassin versant de l'Agly par la méthode hydrogéomorphologique, Technical report, available at: http://piece-jointe-carto.developpement-durable.gouv.fr/REG091B/RISQUE/CDROM/agly/fichiers/rapportAZI_Agly.pdf (last access: 20 November 2019), 2008.

Douinot, A., Roux, H., and Dartus, D.: Modelling errors calculation adapted to rainfall-runoff model user expectations and discharge data uncertainties, Environ. Model. Softw., 90, 157–166, https://doi.org/10.1016/j.envsoft.2017.01.007, 2017.

Douinot, A., Roux, H., Garambois, P.-A., and Dartus, D.: Using a multi-hypothesis framework to improve the understanding of flow dynamics during flash floods, Hydrol. Earth Syst. Sci., 22, 5317–5340, https://doi.org/10.5194/hess-22-5317-2018, 2018.

Drobinski, P., Ducrocq, V., Alpert, P., Anagnostou, E., Béranger, K., Borga, M., Braud, I., Chanzy, A., Davolio, S., Delrieu, G., Estournel, C., Filali Boubrahmi, N., Font, J., Grubišić, V., Gualdi, S., Homar, V., Ivančan-Picek, B., Kottmeier, C., Kotroni, V., Lagouvardos, K., Lionello, P., Llasat, M. C., Ludwig, W., Lutoff, C., Mariotti, A., Richard, E., Romero, R., Rotunno, R., Roussot, O., Ruin, I., Somot, S., Taupier-Letage, I., Tintore, J., Uijlenhoet, R., and Wernli, H.: HyMeX A 10-year multidisciplinary program on the Mediterranean water cycle, B. Am. Meteorol. Soc., 95, 1063–1082, https://doi.org/10.1175/BAMS-D-12-00242.1, 2014.

Dudhia, J.: Numerical study of convection observed during the Winter Monsoon Experiment using a mesoscale two-dimensional model, J. Atmos. Sci., 46, 3077–3107, 1989.

Edouard, S., Vincendon, B., and Ducrocq, V.: Ensemble-based flash flood modelling: Taking into account hydrodynamic parameters and initial soil moisture uncertainties, J. Hydrol., 560, 480–494, https://doi.org/10.1016/j.jhydrol.2017.04.048, 2018.

Evans, J. P., Ekström, M., and Ji, F.: Evaluating the performance of a WRF physics ensemble over south-east Australia, Clim. Dynam., 39, 1241–1258, https://doi.org/10.1007/s00382-011-1244-5, 2012.

Fiori, E., Comellas, A., Molini, L., Rebora, N., Siccardi, F., Gochis, D. J., Tanelli, S., and Parodi, A.: Analysis and hindcast simulations of an extreme rainfall event in the Mediterranean area: the Genoa 2011 case, Atmos. Res., 138, 13–29, 2014.

Fread, D. L.: Flow routing, in: Hanbook of Hydrology, edited by: Maidment, D. R., MCGraw-Hill, Inc., USA, 1992.

Garambois, P. A., Roux, H., Larnier, K., Castaings, W., and Dartus, D.: Characterization of process-oriented hydrologic model behavior with temporal sensitivity analysis for flash floods in Mediterranean catchments, Hydrol. Earth Syst. Sci., 17, 2305–2322, https://doi.org/10.5194/hess-17-2305-2013, 2013.

Garambois, P. A., Roux, H., Larnier, K., Labat, D., and Dartus, D.: Characterization of catchment behavior and rainfall selection for flash flood hydrological model calibration: catchments of the eastern Pyrenees, Hydrolog. Sci. J., 60, 424–447, https://doi.org/10.1080/02626667.2014.909596, 2015a.

Garambois, P.-A., Roux, H., Larnier, K., Labat, D., and Dartus, D.: Parameter regionalization for a process oriented distributed model dedicated to flash floods, J. Hydrol., 525, 383–399, https://doi.org/10.1016/j.jhydrol.2015.03.052, 2015b.

Gaume, E., Bain, V., Bernardara, P., Newinger, O., Barbuc, M., Bateman, A., Blaškovicová, L., Blöschl, G., Borga, M., Dumitrescu, A., Daliakopoulos, I., Garcia, J., Irimescu, A., Kohnova, S., Koutroulis, A., Marchi, L., Matreata, S., Medina, V., Preciso, E., Sempere-Torres, D., Stancalie, G., Szolgay, J., Tsanis, I., Velasco, D., and Viglione, A.: A compilation of data on european flash floods, J. Hydrol., 367, 70–78, https://doi.org/10.1016/j.jhydrol.2008.12.028, 2009.

Gilmour, I., Smith, L. A., and Buizza, R.: Linear region duration: Is 24 hours a long time in synoptic weather forecasting?, J. Atmos. Sci., 58, 3525–3539, 2001.

Green, W. H. and Ampt, C. A.: Studies on soil physics of flow of air and water through soils, J. Agricult. Sci., 4, 1–24, 1911.

Grimit, E. P. and Mass, C. F.: Measuring the ensemble spread–error relationship with a probabilistic approach: stochastic ensemble results, Mon. Weather Rev., 135, 203–221, 2007.

Habets, F., Boone, A., Champeaux, J.-L., Etchevers, P., Franchisteguy, L., Leblois, E., Ledoux, E., Le Moigne, P., Martin, E., Morel, S., Noilhan, J., Quintana Seguí, P., Rousset Regimbeau, F., and Viennot, P.: The SAFRAN-ISBA-MODCOU hydrometeorological model applied over France, J. Geophys. Res.-Atmos., 113, D06113, https://doi.org/10.1029/2007JD008548, 2008.

Hersbach, H.: Decomposition of the Continuous Ranked Probability Score for Ensemble Prediction Systems, Weather Forecast., 15, 559–570, https://doi.org/10.1175/1520-0434(2000)015<0559:DOTCRP>2.0.CO;2, 2000.

Hong, S.-Y. and Lim, J.-O. J.: The WRF single-moment 6-class microphysics scheme (WSM6), J. Korean Meteorol. Soc., 42, 129–151, 2006.

Hong, S.-Y., Noh, Y., and Dudhia, J.: A new vertical diffusion package with an explicit treatment of entrainment processes, Mon. Weather Rev., 134, 2318–2341, https://doi.org/10.1175/MWR3199.1, 2006.

Hsiao, L.-F., Yang, M.-J., Lee, C.-S., Kuo, H.-C., Shih, D.-S., Tsai, C.-C., Wang, C.-J., Chang, L.-Y., Chen, D. Y.-C., Feng, L., Hong, J.-S., Fong, C.-T., Chen, D.-S., Yeh, T.-C., Huang, C.-Y., Guo, W.-D., and Lin, G.-F.: Ensemble forecasting of typhoon rainfall and floods over a mountainous watershed in Taiwan, J. Hydrol., 506, 55–68, https://doi.org/10.1016/j.jhydrol.2013.08.046, 2013.

Hu, X.-M., Nielsen-Gammon, J. W., and Zhang, F.: Evaluation of three planetary boundary layer schemes in the WRF Model, J. Appl. Meteorol. Clim., 49, 1831–1843, https://doi.org/10.1175/2010JAMC2432.1, 2010.

Jain, S. K., Mani, P., Jain, S. K., Prakash, P., Singh, V. P., Tullos, D., Kumar, S., Agarwal, S. P., and Dimri, A. P.: A Brief review of flood forecasting techniques and their applications, Int. J. River Basin Manage., 16, 329–344, https://doi.org/10.1080/15715124.2017.1411920, 2018.

Janjic, Z. I.: The step-mountain eta coordinate model: further developments of the convection, viscous sublayer, and turbulence closure schemes, Mon. Weather Rev., 122, 927–945, https://doi.org/10.1175/1520-0493(1994)122<0927:TSMECM>2.0.CO;2, 1994.

Jankov, I., Gallus Jr., W. A., Segal, M., Shaw, B., and Koch, S. E.: The impact of different WRF Model physical parameterizations and their interactions on warm season MCS rainfall, Weather Forecast., 20, 1048–1060, https://doi.org/10.1175/WAF888.1, 2005.

Krzysztofowicz, R.: The case for probabilistic forecasting in hydrology, J. Hydrol., 249, 2–9, https://doi.org/10.1016/S0022-1694(01)00420-6, 2001.

Ladouche, B. and Dörfliger, N.: Evaluation des ressources en eau des corbières. Phase I – Synthèse de la caractérisation des systèmes karstiques des Corbières Orientales, Tecnical report BRGM, available at: http://infoterre.brgm.fr/rapports/RP-52919-FR.pdf (last access: 6 December 2019), 2004.

Laurantin, O.: ANTILOPE: hourly rainfall analysis merging radar and raingauges data, in: Proceedings of Weather Radar and Hydrology Conference 2008, Grenoble, 2008.

Le Lay, M. and Saulnier, G. M.: Exploring the signature of climate and landscape spatial variabilities in flash flood events: case of the 8–9 September 2002 Cévennes-Vivarais catastrophic event, Geophys. Res. Lett., 34, L13401, https://doi.org/10.1029/2007GL029746, 2007.

Leoncini, G., Plant, R. S., Gray, S. L., and Clark, P. A.: Ensemble forecasts of a flood producing storm: comparison of the influence of model-state perturbations and parameter modifications, Q. J. Roy. Meteorol. Soc., 139, 198–211, 2013.

Mansell, E. R.: On sedimentation and advection in multimoment bulk microphysics, J. Atmos. Sci., 67, 3084–3094, https://doi.org/10.1175/2010JAS3341.1, 2010.

Matheson, J. E. and Winkler, R. L.: Scoring rules for continuous probability distribution, Manage. Sci., 22, 1087–1096, 1976.

Maubourguet, M.-M., Chorda, J., Dartus, D., and George, J.: Prévision des crues éclair sur le Gardon d'Anduze (Flash flood forecasting in the Gardon catchment at Anduze), in: 1st Mediterranean-HyMeX Workshop – Hydrological cycle in Mediterranean Experiment, 9–11 January 2007, Météo-France, Toulouse, France, 2007.

Mlawer, E. J., Taubman, S. J., Brown, P. D., Iacono, M. J., and Clough, S. A.: Radiative transfer for inhomogeneous atmospheres: RRTM, a validated correlated-k model for the longwave, J. Geophys. Res., 102, 16663–16682, 1997.

Molteni, F., Buizza, R., Palmer, T. N., and Petroliagis, T.: The ECMWF ensemble prediction system: methodology and validation, Q. J. Roy. Meteorol. Soc., 122, 73–119, https://doi.org/10.1002/qj.49712252905, 1996.

Mounier, F., Lassègues, P., Gibelin, A.-L., Céron, J.-P., and Veysseire, J.-M.: Radar-guided control and interpolation of rain gauge precipitation data over France, Report EURO4M project (European Reanalysis and Observations for Monitoring project), available at: http://www.euro4m.eu/Publications/Report_Flore_Mounier_EURO4M_201203.pdf (last access: 6 December 2019), 2012.

Nakanishi, M. and Niino, H.: An Improved Mellor–Yamada Level-3 Model: Its Numerical Stability and Application to a Regional Prediction of Advection Fog, Bound.-Lay. Meteorol., 119, 397–407, https://doi.org/10.1007/s10546-005-9030-8, 2006.

Nash, J. E. and Sutcliffe, J. V.: River flow forecasting through conceptual models. Part I – A discussion of principles, J. Hydrol., 10, 282–290, https://doi.org/10.1016/0022-1694(70)90255-6, 1970.

Nurmi P.: Recommendations on the verification of local weather forecasts, ECMWF Tech. Mem. 430, available at: https://www.researchgate.net/publication/238107438_Recommendations_on_the_verification_of_local_weather_forecasts (last access: 20 November 2019), 2003.

Pilgrim, D. H. and Cordery, I.: Flood runoff, in: Hanbook of Hydrology, editedd by: Maidment, D. R., MCGraw-Hill, Inc., USA, 1992.

Ramos, M. H., van Andel, S. J., and Pappenberger, F.: Do probabilistic forecasts lead to better decisions?, Hydrol. Earth Syst. Sci., 17, 2219–2232, https://doi.org/10.5194/hess-17-2219-2013, 2013.

Ravazzani, G., Amengual, A., Ceppi, A., Homar, V., Romero, R., Lombardi, G., and Mancini, M.: Potentialities of ensemble strategies for flood forecasting over the Milano urban area, J. Hydrol., 539, 237–253, https://doi.org/10.1016/j.jhydrol.2016.05.023, 2016.

Rossa, A. M., Laudanna Del Guerra, F., Borga, M., Zanon, F., Settin, T., and Leuenberger, D.: Radar-driven high-resolution hydro-meteorological forecasts of the 26 September 2007 Venice flash flood, J. Hydrol., 394, 230–244, https://doi.org/10.1016/j.jhydrol.2010.08.035, 2010.

Roux, H., Labat, D., Garambois, P.-A., Maubourguet, M.-M., Chorda, J., and Dartus, D.: A physically-based parsimonious hydrological model for flash floods in Mediterranean catchments, Nat. Hazards Earth Syst. Sci., 11, 2567–2582, https://doi.org/10.5194/nhess-11-2567-2011, 2011.

Salvayre, H.: Les karsts des Pyrénées-Orientales (Caractères hydrogéologiques et spéléologiques généraux), in: Karstologia: revue de karstologie et de spéléologie physique, no. 13, 1er semestre 1989, 1–10, https://doi.org/10.3406/karst.1989.2199, 1989.

Schwartz, C. S., Kain, J. S., Weiss, S. J., Xue, M., Bright, D. R., Kong, F., Thomas, K. W., Levit, J. J., Coniglio, M. C., and Wandishin, M. S.: Toward improved convection-allowing ensembles: Model physics sensitivities and optimizing probabilistic guidance with small ensemble membership, Weather Forecast., 25, 263–280, https://doi.org/10.1175/2009WAF2222267.1, 2010.

Skamarock, W. C., Klemp, J. B., Dudhia, J., Gill, D. O., Barker, D., Duda, M. G., Huang, X. Y., Wang, W., and Powers, J. G.: A Description of the Advanced Research WRF Version 3, NCAR Tech. Note NCAR/TN-4751STR, NCAR, Boulder, Colorado, USA, p. 125, 2008.

Siddique, R. and Mejia, A.: Ensemble Streamflow Forecasting across the U.S. Mid-Atlantic Region with a Distributed Hydrological Model Forced by GEFS Reforecasts, B. Am. Meteorol. Soc., 18, 1905–1928, https://doi.org/10.1175/JHM-D-16-0243.1, 2017.

Stensrud, D.J., Bao, J.-W., and Warner, T. T.: Using initial and model physics perturbations in short-range ensemble simulations of mesoscale convective events, Mon. Weather Rev., 128, 2077–2107, 2000.

Tao, W. K., Simpson, J., and McCumber, M.: An ice-water saturation adjustment, Mon. Weather

Rev., 117, 231–235, https://doi.org/10.1175/1520-0493(1989)117<0231:AIWSA>2.0.CO;2, 1989.

Tapiador, F. J., Tao, W. K., Shi, J. J., Angelis, C. F., Martinez, M. A., Marcos, C., Rodríguez, A., and Hou, A.: A comparison of perturbed initial conditions and multiphysics ensembles in a severe weather episode in Spain, J. Appl. Meteorol. Clim., 51, 489–504, 2012.

Tewari, M., Chen, F., Wang, W., Dudhia, J., LeMone, M. A., Mitchell, K., Ek, M., Gayno, G., Wegiel, J., and Cuenca, R. H.: Implementation and verification of the unified NOAH land surface model in the WRF model, in: 20th Conference on Weather Analysis and Forecasting/16th Conference on Numerical Weather Prediction, 11–15 January 2004, Seattle, Washington, 11–15, 2004

Thiessen, A. H.: Precipitation averages for large areas, Mon. Weather Rev., 39, 1082, https://doi.org/10.1175/1520-0493(1911)39<1082b:PAFLA>2.0.CO;2, 1911.

Thompson, G., Field, P. R., Rasmussen, R. M., and Hall, W. D.: Explicit forecasts of winter precipitation using an improved bulk microphysics scheme. Part II: Implementation of a new snow parameterization, Mon. Weather Rev., 136, 5095–5115, https://doi.org/10.1175/2008MWR2387.1, 2008.

Todini, E.: Role and treatment of uncertainty in real-time flood forecasting, Hydrol. Process., 18, 2743–2746, https://doi.org/10.1002/hyp.5687, 2004.

Verkade, J. S., Brown, J. D., Reggiani, P., and Weerts, A. H.: Post-processing ECMWF precipitation and temperature ensemble reforecasts for operational hydrologic forecasting at various spatial scales, J. Hydrol., 501, 73–91, https://doi.org/10.1016/j.jhydrol.2013.07.039, 2013.

Verkade, J. S., Brown, J. D., Davids, F., Reggiani, P., and Weerts, A. H.: Estimating predictive hydrological uncertainty by dressing deterministic and ensemble forecasts; a comparison, with application to Meuse and Rhine, J. Hydrol., 555, 257–277, https://doi.org/10.1016/j.jhydrol.2017.10.024, 2017.

Zappa, M., Beven, K. J., Bruen, M., Cofiño, A. S., Kok, K., Martin, E., Nurmi, P., Orfila, B., Roulin, E., Schröter, K., Seed, A., Szturc, J., Vehviläinen, B., Germann, U., and Rossa, A.: Propagation of uncertainty from observing systems and NWP into hydrological models: COST-731 Working Group 2, Atmos. Sci. Lett., 11, 83–91, https://doi.org/10.1002/asl.248, 2010.

Zappa, M., Jaun, S., Germann, U., Walser, A., and Fundel, F.: Superposition of three sources of uncertainties in operational flood forecasting chains, Atmos. Res., 100, 246–262, https://doi.org/10.1016/j.atmosres.2010.12.005, 2011.

Zappa, M., Fundel, F., and Jaun, S.: A 'Peak-Box' approach for supporting interpretation and verification of operational ensemble peak-flow forecasts, Hydrol. Process., 27, 117–131, https://doi.org/10.1002/hyp.9521, 2013.