Response to Reviewer 2

December 2019

1 General comment

For this study the authors put a lot of effort in analyzing an operational used weather forecast model with respect to its performance and applicability on heavy precipita- tion events (HPE) in the Mediterranean region. The authors create a 30-year long 10 member hindcast ensemble using different parameterization schemes for convection and compared simulated HPEs with observations. HPEs are of great importance for that region as they are relatively frequent in the autumn and early winter season. Se- vere flooding and damaging are related hazards. This study falls within in the scope of NHESS. The title of this study sounds very promising in giving some real benefits to improve the performance of numerical models in predicting extreme events. Unfortunately, this is not the case in my opinion and I miss the added value of this study. My fundamental concern with this study is the chosen model. The authors used PEARP, an ensemble using the global model of the French Weather Service ARPEGE. Even though it has an in-model nesting of different grids down to a highest horizontal resolution of 10 km over France, convection is parameterized using known convection schemes as described in the data section 2. But, deep moist convection generates most of the precipitation amounts during HPEs in the western Mediterranean. The global model with parameterized convection is not meant to simulate such events properly. I would have expected an analysis of prediction errors using a higher resolved regional and convection permitting model like AROME or ALADIN, both also run operational by the French Weather Service. Therefore, the authors' conclusions, e.g. that the size of simulated object is larger than in observations but the amplitude is reduced, seem to me quiet obvious and more a consequence of the parameterization, which is already known and nothing new.

Beside my maybe wrong expectation, I do see the point, that the coarser model is cheaper in computation time and therefore it is worth looking at systematic errors, but as there is a trend to more and more higher resolutions for weather forecasts it should be stated clearly what the benefits of the coarse model would be. Nevertheless, the presented methodology is interesting and suitable for such kind of study. Furthermore, analyzing possible systematic errors especially in predicting extremes is also very im- portant and improvements would give benefit to different applications. Beside my main concern above, I

have a few major comments and some specific points listed below.

RESPONSE: Thanks to the referee for this global comment. We agree that we have to clarify in the paper the reasons of the use of a global model rather than a convection-permitting high-resolution one. The first reason is, as stated by the referee, the numerical cost of a 30-year high-resolution reforecast. At the moment, such a tool (based for example on the AROME model) does not exist at Météo-France and the numerical cost involved in its building would have been too high for the study. The second reason is that one goal of the study was to explore systematic forecast errors up to 4-day lead-time. Although we agree that high resolution models (eg 2.5 km for the Météo-France AROME-EPS) are of primary interest for very short lead-time forecast (e.g less than 48h lead-time) we think that small-scale predictability is roughly lost beyond 2-day lead-time. Using a global operational 10-km EPS allows to explore HPE predictability up to several days. We also agree with the referee that the fact that the size of simulated rainfall objects is larger than in observations and their amplitude is reduced is not surprising and may be mainly a consequence of the parameterization. However we would like to point out that the conclusions of our paper are based on the study of around 200 cases of HPE, over a 30-year period, and include verification for very high rainfall thresholds. Most of previous papers on the subject focus on one or a few iconic cases. Here, using reforecast allows to hold this systematic review of French HPEs over the last 30 years and the opportunity to use very high thresholds for verification (something that cannot be done in operational verification due to the shortness of the verification periods). Some sentences have been added to the introduction and the conclusion to explain more clearly the aims and conclusions of the paper. They are detailed below.

2 Major comments

- The paper is hard to read due to some language deficits especially when it comes to the technical parts. I would strongly recommend a revision on sentence structure, grammar, comma, or word usage.

  RESPONSE: A professional proofreading to correct English language deficits has been performed.

- The authors only analyzed precipitation fields and differences between the param- eterization schemes for deep convection using the SAL method. A broader look on other quantities like ambient and/or convection favoring conditions is missing. Initial and boundary conditions as well as model physics related to the model resolution have a significant influence on the simulation of convection as presented, for example, in Kunz et al, 2018 (doi: 10.1002/qj.3197), Khodayar et al., 2018 (doi: 10.1002/qj.3402) or Caldas-Álvarez et al., 2019 (doi: 10.5194/asr-14-157-2017). Furthermore,

local dy- namic pattern also influence the initialization of convection especially in mountainous terrain or on islands (e.g. Ehmele et al, 2015; doi: 10.1016/j.atmosres.2014.10.004), so a misrepresentation of these also lead to distinct differences between model and observations. A third thing are specific weather patterns which have an influence on ambient conditions and convection. Errors or deviations in the model regarding such patterns will also cause a bad representation of HPEs as well. A connection of weather patterns to convection across Central Europe (including France) can be found, for in- stance, in Piper et al., 2019 (doi: 10.1002/qj.3647).

RESPONSE: We agree with the referee that many factors can be considered as sources of HPEs forecast errors such as uncertainties in the initial state of the forecast, the synoptic-scale configuration, the local dynamical effects, etc. Here our goal was not to analyse the HPEs forecast errors through all their potential sources but to focus on errors that comes from the parameterizations of the main physical processes (for initial state errors, it is assume that, as, for a given day, all forecast of the reforecast have the same initial state, initial state uncertainties will have the same impact on each of the 10 forecasts) The use of a multi-physics approach is classical in ensemble forecasting and one main goal of our study was to evaluate this approach through a systematic analysis of forecast errors of each set of physical parameterizations. This is also the reason why we focused on rainfall field as we assume that misrepresentation of processes in the parameterizations or bad combinations of schemes in the multi-physics will finally produce forecast errors in the rainfall field.

- What is the added value of this study? This is the crucial thing of this study and should be strongly pointed out not only but especially in the conclusions section. Additionally, some concrete statements on how to apply the results in terms of future model improvements should be given so that the reader can really benefit from this study.

  RESPONSE: We agree with the referee that the paper does not enough point out the main benefits of the study. The main goals, conclusions and benefits of the study have been emphasized in the revised manuscript. As an example, we have mentioned in the conclusion that this forecast analysis gives practical information to modellers as well as to forecasters. To modellers of ensemble prediction systems, the study clearly shows the limits of the multiphysics approach in the representation of model errors. Indeed the study shows that 'the real variability' of such an approach could be limited to only a very few 'actual different behaviours' of the different physics. This conclusion is also important for forecasters who use operational system based on multiphysics approach and all the information on forecast errors extracted from the study could help them to better understand the behaviour of the operational system. The conclusion has been re-organised and expanded with clearer points about the results of

the study.

CHANGE: Specific statements have been proposed through some modifications in the whole manuscript. Here we report the major modifications concerning the Abstract, Introduction and Conclusion sections.

Abstract: text from [1 4] to [1 11] is replace by:
"In this study, we analyze the HPE forecasting ability of the multi-physics based ensemble model operational at Météo-France Prévision d'Ensemble ARPEGE (PEARP). The analysis is based on 30-year (1981-2010) ensemble hindcasts which implement the same 10 physical parametrizations, one per member, run every 4 days. Over the same period a 24-hour precipitation dataset is used as the reference for the verification procedure. Furthermore, regional classification is performed in order to investigate the local variation of spatial properties and intensities of rainfall fields, with a particular focus on HPEs. As gridpoint verification tends to be perturbed by the double penalty issue, we focus on rainfall spatial pattern verification thanks to the feature-based quality measure SAL that is performed on the model forecast and reference rainfall fields. The length of the dataset allows to sub-sample scores for very intense rainfall at a regional scale and still get significant analysis demonstrating that such a procedure is consistent to study model behaviour in HPE forecasting. In the case of PEARP, we show that the amplitude and structure of the rainfall patterns are basically driven by the deep convection parametrization."

Introduction: from [2 31] to the end:
"Ehmele et al. (2015) emphasized the important role played by complex orography, the mutual interaction between two close mountainous islands in this case, on heavy rainfall under strong synoptic forcing conditions. Nevertheless, other regions are also affected by rainfall events with a great variety of intensity and spatial extension. Ricard et al. (2011) studied this regional spatial distribution based on a composite analysis and showed the existence of mesoscale environments associated with heavy precipitating events. Considering four sub-domains, they found that the synoptic and mesoscale patterns can greatly differ as a function of the location of the precipitation.

Extreme rainfall events are generally associated with coherent structures slowed down and enhanced by the relief, whose extension is often larger than a single thunderstorm cell. At some point, this mesoscale organization can turn into a self-organization process leading to a mesoscale convective system (MCS) when interacting with their environment, which in turn leads to high intensity rainfall (Nuissier et al., 2008).

Among the list of factors contributing to HPE creation, some are clearly only within the scope of high resolution convection permitting models. Indeed, vertical motion and moisture processes need to be explicitly solved

to get realistic representation of convection. On the other hand, as we have just highlighted, some other factors linked with synoptic circulations or orography representations can be well estimated in global models, in particular when horizontal resolution gets close to 15-20 km. Consequently, the corresponding predictability of such factors can reach advantageous lead times for early warnings, i.e. longer than the standard 48 hours that the limited area model may be expected to achieve. Indeed, if long term territorial adaptations are necessary to mitigate the impact of HPEs, a more reliable and earlier alert would be beneficial in the short term. Weather forecasting coupled with hydrological impact forecasting is the main source of information for triggering of weather warnings. Severe weather warnings are issued for the 24-hour forecast only. However, in some cases, the forecast process could be issued some days prior to the severe weather warnings. A better understanding of the sources of model uncertainty at such time-range may provide a major source of improvement for early diagnosis.

Forecast uncertainties can be related to initialization data (analysis) or lateral boundary conditions, and it has been investigated with both deterministic models (Argence et al., 2008) and ensemble models (Vié et al., 2010). Several journal articles showed that predictability associated with intense rainfall and flash-floods decreases rapidly with the event scale (Walser et al., 2004; Walser and Schär, 2004; Collier, 2007). Several studies based on ensemble prediction systems have shown that such models may have a great ability in sampling the sources of uncertainty in HPE probabilistic forecasting (Du et al., 1997; Petroliagis et al., 1997; Stensrud et al., 1999; Schumacher and Davis, 2010; World Meteorological Organization, 2012). In ensemble forecasting, the uncertainty associated with the forecast is usually assessed by taking into account initial and model error propagation. As for the initial uncertainty, major meteorological centers implement different methods the most common of which are singular vectors (Buizza and Palmer, 1995; Molteni et al., 1996) , bred vectors (Toth and Kalnay, 1993, 1997) and perturbed observation in analysis process (Houtekamer et al., 1996; Houtekamer and Mitchell, 1998). The model error is related to grid-scale unsolved processes in the parametrization scheme and is assessed in the models with two main techniques. Some models use stochastic perturbations of the inner-model physics scheme (Palmer et al., 2009), others use different parametrization schemes in each forecast member (Charron et al., 2009; Descamps et al., 2011).

The global ensemble model implemented at Météo-France Prévision d'Ensemble ARPEGE (PEARP; Descamps et al., 2015) is based on the second technique, also known as a multi-physics approach. Compared to the stochastic perturbation, the error model distribution cannot be explicitly formulated in the multi-physics approach. It is then difficult to know a priori the influence of the physics scheme modifications on the forecast ability of the model. This is even more the case when highly non-linear physics

with high order of magnitude processes are considered. In order to improve the understanding and interpretation of ensemble forecasts in tense decision-making situations as well as for model development and improvement purposes, it would be of great interest to have a full and objective analysis of the model behaviour in terms of HPE forecasting. This is one of the main aims of this study.

In order to achieve such a systematic analysis, standard rainfall verification methods can be used. They are usually based on grid-point based approaches. These techniques, especially when applied to intense events, are subject to time or position errors leading to low scores (Mass et al., 2002) also known as the double penalty problem (Rossa et al., 2008). To counteract this problem, spatial verification techniques have been developed with the goal of evaluating a forecast quality from a forecaster standpoint. Some of these techniques are based on object-oriented verification methods (AghaKouchak et al., 2011; Ebert and McBride, 2000; Davis et al., 2006, 2009; Mittermaier et al., 2015; Wernli et al., 2008).The feature-based quality measure SAL (Wernli et al., 2008, 2009) is also used in this study. Another element required to achieve such an analysis is the availability of forecast datasets long enough to get a proper sampling of the events to verify.

In our study, we profit from a reforecast dataset based on a simplified version of the PEARP model available over a 30 year period. Such reforecast datasets have been previously shown to be relevant for calibrating operational models in various ways. In (Hamill and Whitaker, 2006; Hamill et al., 2008; Hamill, 2012; Boisserie et al., 2015), the reforecast is used as a learning dataset to fit statistical models to calibrate forecast error corrections that are then applied on operational forecasting outputs. (Boisserie et al., 2015; Lalaurette, 2003) have shown the possibility of using a reforecast dataset as a statistical reference of the model to which the extremeness of a given forecast is compared. In this paper, we analyze the ensemble model PEARP forecast predictability at lead times between day 2 and day 4 of daily rainfall amounts. This analysis is performed on the long reforecast 30-year dataset. One aim is to determine whether a multi-physics approach could be considered as a model error sampling technique appropriate for a good representation of HPEs in the forecast at such lead times. In particular, the behaviour of the different physics schemes implemented in PEARP have to be estimated individually. One main side aspect of this work focuses on developing a methodology suitable for evaluating the performances of an ensemble reforecast in a context of intense precipitation events using an object oriented approach. In particular we focus on intense precipitation over the French Mediterranean region. In addition to the analysis of diagnostics from the SAL-metric, a statistical analysis of 24-hour rainfall objects identified in the forecasts and the observations is performed in order to explore the spatial properties of the rainfall fields.

The data and the methodology are presented in section 2. Section 2.1 describes the reforecast ensemble dataset and section 2.2 details the creation of the daily rainfall reference, the HPEs statistical definition and the regional clustering analysis. Results arising from the spatial verification of the overall reforecast dataset are presented in section 3.1. Section 3.2 presents SAL diagnostics divided into all different physical parametrization schemes of the ensemble reforecast, and for the spatial properties of individual objects. Conclusions are given in section 4."

Conclusion section from [28 493] to [30 526] is replaced by:
"The peak-over-threshold criterion leads to the selection of 192 HPEs, confirming that the most impacted region are the Cévennes area and part of the Alps. The composite analysis for the five clusters shows that each cluster is associated with a specific class and location of 24-hour precipitation events. It was found that 86% of the number of HPEs are included in clusters 2, 3 and 5. Cluster 2 and 3 HPEs predominantly impact the Cévennes and Alps area, while cluster 5 HPEs are located over the Languedoc-Roussillon region. Moreover clusters 3 and 5 include the most extreme ones. Only diagnostics for clusters 2, 3 and 5 are considered.

The SAL object-quality measure has been applied distinctly to the ten physics schemes (one per member) of the reforecast dataset and compared to the rainfall reference. It shows that the model's overall behaviour for HPE forecasting is characterized by negative A-components and positive S-components. As in grid-point rainfall verification, all the SAL components get worse as a function of lead time. Then the model HPE rainfall objects tend to be more extended and less peaked. Even though their corresponding domain-average amplitude is weaker, it doesn't mean that the event maximum intensity is always weaker. This result is important showing to modelers that even for intense rainfall events when orography interaction and quasi-stationarity meso-scale systems play a great role, the model tends to reproduce rainfall patterns with greater extension, rather than both smaller extension and weaker intensity patterns.

In order to show regional disparities in the model behaviour, the SAL diagnostics have been divided according to the clusters and it shows interesting results. First, the A component negative contribution for the whole sample is higher, showing that in average more underestimation than overestimation is observed for the Aplitude SAL-component. It is notably the case for the most extreme clusters (over the Cévennes and over the Languedoc-Roussillon). However, when considering both positive and negative contributions to the integrated A-component, the most extreme cluster (cluster 3) leads to better scores. This could mean that the variability of the A-component is postively reduced for the most intense events. This is quite surprising and could reinforce the role of orography in this error decrease. As for the S-component distribution, we showed it is slightly positively skewed for cluster 2 and 3, while for cluster 5 the

distribution of the S-component is more centered. Likewise for the A-component the integrated balance of positive and negative S-component contributions lead to better results for cluster 3 and 5. It is even more remarkable for cluster 5, for which the S-component reaches the best score. Though though it is difficult at this point to determine whether this characterizes an actual contrast in the model behaviour or if it is due to the physical properties of the cluster 5 events. One hypothesis could be related to the large number of single objects characterizing this cluster.

The impact of the different physics schemes has also been investigated, moThe impact of the different physics schemes has also been investigated, and it mostly emphasized the role of the deep convection physical parameterization. Considering the SAL diagnostics, the two main deep convection schemes, B85 and PCMT, clearly determine the behaviour of the model in HPE forecasting until lead time ranges longer than three days, after which no significant differences appear. This difference is clearly in favour of the PCMT scheme which performs better than B85 for both SAL A and S components and in the majority of the subsampled scores considering the HPEs or the regional clusters. However, this PCMT asset is not huge, and both physics schemes can contribute to good or bad forecasts. The main significant difference is for the S-component for the most intense rainfall, which shows that PCMT better approximates the structure of the rainfall patterns in these cases.

In light of the ability of our method to produce significant results even after several subsampling steps, we decided to study further statistical characterization of the SAL rainfall objects. It has been shown that in most cases, one large object stands out among other smaller objects, which often gathers the most part of the rain signal. For cluster 5, characterized by the Languedoc-Roussillon HPEs, the rainfall distribution could even be considered as a single object rainfall field. Then we focused on the ranked distributions (quantile-quantile analysis) of the object masses to compare the rainfall model overall climatology of the model with the reference. First, this analysis showed that in particular the weakest precipitation are overestimated by all physics schemes. However, looking at the object mass distributions for the whole period, we find they are relatively close between all the physics schemes and the observation for most extreme rainfall events, especially for the PCMT deep convection scheme. This statistical result implies that a global model should be able to reproduce a reliable distribution of rainfall objects along a long time period, e.g. the climate of the model and of the observations are close to each other. Therefore, in the case of PEARP, most of the forecast errors are mainly related to a low consistency between observed and forecasted fields, rather than to an inability of the prediction system to produce intense precipitation amounts.

This last result, objectively quantified for high rainfall event thresholds (around 100 mm to 500 mm) on a long enough period, is important for

two reasons. The first concerns atmospheric modelers, showing that the physics schemes are able to reproduce climatological distributions of the most challenging rainfall events. On this basis, future research could investigate other sources of uncertainties like from the analysis setup and implement ensuing model improvements. The model physics perturbation technique should then play a greater role in the control of the ensemble dispersion. In this perspective, the novel reanalysis ERA5 would be interesting to use, in particular its perturbed members, to improve the uncertainty from initial conditions in the reforecast. The second lesson to be learned from this study is that it is worth spending time to study a model behaviour on intense events forecasting as it provides important learning to ensemble model end-users, in particular in the context of decision making based on weather forecast. Quantifying systematic errors could also be used to favorably improve their inclusion in nested forecast tools processes.

In terms of methodology, this study also highlights that the combination of SAL verification and clustering is a relevant approach to show systematic errors associated with regional features for intense precipitation forecasting. This achievement is only enabled by the availability of a long reforecast dataset. This methodology could be further extended to a different model and another geographic region, on the condition of sampling a large number of HPEs. ”

In order to reply to the recommendation of referee Number 1, some shortening has been done:
Technical description of k-sample Anderson-Darling test has been removed. Lines from [18 367] to [18 368]. Eq. (9) and (10) are removed. Tables 5 and 6 are removed. Tables 7,8,9, and 10 have been replaced by a new figure. Fig. 12 and the corresponding text is removed.

3 Minor comments

1 18 ’[...] daily rainfall amounts associated to a one single event’, ’a single event’ or ’one single event’

RESPONSE: The text has be corrected

CHANGE: “a single event”

2 27 ’4) a synoptic system to slow the convective system [...]’, I think you mean ’to hold’ or better ’to retain’

RESPONSE: The text has be corrected

CHANGE: “a quasi-stationary convective system that persists over the threat area”

2 30ff Another study analyzing extreme precipitation in the Mediterranean, also both pure convective and convection-stratiform mix, and related mechanisms and processes is presented in Ehmele et al., 2015 (doi: 10.1016/j.atmosres.2014.10.004).

RESPONSE: Thanks to the referee for this bibliographic suggestion. This reference has been added to the Introduction.

CHANGE: [2 31] the following sentence is added "Ehmele et al. (2015) emphasized the important role played by complex orography, the mutual interaction between two close mountainous islands in this case, on heavy rainfall under strong synoptic forcing conditions"

3 88 'affected by the precipitation', not 'precipitations'

RESPONSE: The text has been corrected

CHANGE: As suggested by the reviewer

- Figure 1: speaking of domain D, it should be given in the plot where D exactly is. Is it the red box in (a) meaning the whole plot area of (b) and (c)?

  RESPONSE: Thanks to the referee for this suggestion to give better specifications about the domain. The domain D corresponds solely to the model grid shown in blue in (c)

  CHANGE: In Figure 1: "Panel **c** shows the $0.1° \times 0.1°$ model grid (in blue), along with the location of three key areas. The domain D is located within the borders of the model grid (panel **c**)."

- Table 1: Why only this combinations of parameterization schemes? CAPE is only used for one simulation while B85 is used 5 times or PCMT 3 times, for example TKE + CAPE is missing and so on. Why don't you use equal numbers of every possible combination?

  RESPONSE: Thanks to the referee for this question concerning the combination of physical schemes. In this study we assessed the multiphysics approach implemented in the operational Ensemble Prediction System PEARP. In the context of verification of an operational model, the same physical packages as the ones implemented in PEARP are used. These combinations are developed, tested and maintained by a scientific team at CNRM/Météo-France.

  CHANGE: [112 4] "The same nine different physical parametrizations as the ones used in PEARP (see Table 1) are added to the one corresponding to the ARPEGE deterministic physical package."

7 160ff First you say threshold T = 85mm, but then it is 100mm. So what is the correct threshold you have used? Is it the same threshold or something different? This needs to be clarified. Furthermore, you define a HPE with a single grid point reaches 100mm. You have interpolated to point observations to a regular grid. Is it possible that you miss events due to this interpolation meaning that an exceedance of 100mm at a single grid

is too high? What about HPEs with rainfall amounts below the threshold for 24h but excessive rainfall over 48h or 72h?

RESPONSE: Thanks to the referee for requiring details about the classification of HPEs. The use of different thresholds can be misleading, indeed. First a 85 mm threshold has been selected to split the domain into two regions. Grid observation points where the 99.5 percentile is larger than 85 mm corresponds to regions where intense rainfall commonly occur (sub-region A in the article), while the remaining region (essentially the plain area) tends to be characterised by a lower number of intense rainfall, corresponding to few cases of HPEs (sub-region B in the article). Then, in the sub-region A we applied the 99.5 percentile threshold to identify HPEs, whereas a 100 mm is applied on the sub-region B. In this latter sub-region we preferred to use 100 mm rather than the 99.5 percentile threshold because this latter threshold would be equal to low values (30-40 mm). These precipitation amounts are unlikely associated with HPEs. Second, we agree that the interpolation may have an impact on the HPEs selection, as interpolated values are filtered. Then, it is possible that some events could be missed due to the interpolation procedure. However, we believe that this approach is more proper than a selective method computed over each rain-gauges. The identification of HPEs per grid point assures a spatial homogeneity and a temporal continuity over the 30-year period.
In this study, we focused on daily precipitation, as, e.g., in Ricard et al. (2011), or Ramos et al. (2015). An integration over a longer period (like 48h or 72h) would have reduced the number of cases and available forecasts. On the other hand, the use of precipitation values at a larger frequency would have dramatically reduced the number of available observation rain-gauges.

CHANGE: text from [7 159] to [8 172] is replaced by "We proceed as follows: first the domain is split into two sub-regions based on the occurrence of climatological intense precipitations during the 30 year period. The sub-region A includes all the points whose climatological 99.5 percentile is lower or equal to a threshold T, subregion B includes all the other points. Threshold T, after several tests, has been set to 85 mm. This choice was made in order to separate the domain into two regions characterized by different frequency and intensity of HPEs. Subregion A designates a geographical area where a large number of cases of intense precipitation are observed. Subregion B primarily covers the plain area, where HPE frequency is lower. For this reason, two different level thresholds values are selected to define an event, depending on the subregion. More specifically, a day is classified as an HPE if one point of sub-region A accumulated rainfall is greater than 100 mm or if one point of sub-region B rainfall is greater than its 99.5 percentile. The selection led to a classification of 192 HPEs, corresponding to a climatological frequency of 5% over the 30-year period. The 24-hour rainfall amount maxima within the HPE

dataset ranges from 100 mm to 504 mm. It is worth mentioning that since we consider daily rainfall, rainfall events that would have high 48 hour or 72 hour accumulated rainfall may be disregarded. Figure 2 shows for each point of the domain the number of HPE, as well as the composite analysis of HPEs. The composite analysis involves computing the grid point average from a collection of cases. The signal is enhanced along the Cévennes chain and on the Alpine region. It should be noted that some points are never taken into account for the HPE selection (grey points of Fig. 2), because the required conditions have not been met. The analysis of the rainfall fields across the HPE database exhibits the presence of patterns of different shape and size, revealing potential differences in terms of the associated synoptic and mesoscale phenomena (not shown). "

7 165 so 192 HPE days in 30% is 5%, I agree. The 99.5% percentile would be 18 days in 30 years. Can you please explain the difference?

RESPONSE: Thanks to the referee for suggesting a clearer explanation about the number of identified HPEs. Since the peak-over-threshold approach is separately applied to each point, it is sufficient to observe an exceeding at a given point over the domain to identify an HPE. Similarly, a co-occurence due to the exceeding of thresholds at several grid points at a given day is still considered as one single event at that specific day. As a result, the total number of HPE does not corresponds to 0.005 frequency. It would have been the case if the peak-over-threshold approach had been applied to the whole domain. However, using this approach almost only HPEs impacting the Cévennes area would have been detected, since the most intense events have been observed over this area. This latter evidence explains why a grid-point threshold has been preferred.

- Table 2: I do not understand the difference between HPEs (%) and Fraction of HPEs (%). Can you please specify?

  RESPONSE: Thanks to the referee for this question. This needed to be clarified. A specification is added to the Table 2 labels.

  CHANGE: Table 2: "HPEs(%) refers to the ratio between the number of HPEs within the cluster and the total number of HPEs. Fraction of HPEs (%) refers to the ratio between the number of HPEs within the cluster and the total number of dates included in the corresponding cluster."

9 196 Cluster 5 contains 86% of the HPEs. In Table 2, it says Fraction of HPEs is 65.2%. Should this be the same?

RESPONSE: Thanks to the referee for reporting this mistake about the Fraction of HPEs. We should have state that 86% of HPEs is included in among clusters 2,3 and 5. The text has been corrected.

CHANGE: [9 194] "86%"→"65%", "Clusters 2,3 and 5 collect together 86% of the HPEs."

11 248 Equation (6): I think the 'x element of Obj_k' should not be below the fraction but behind?

RESPONSE: Thanks to the referee for this suggestion. The notation in Equation 6 is modified as suggested.

CHANGE:

$$V_k = \frac{M_k}{\max R(x; x \in Obj_k)}. \quad (1)$$

13 280ff Are there some simulated HPE days among the false alarms?

RESPONSE: Thanks to the referee for this question. It is useful to specify the number of HPEs within the False Alarms, because it would imply that some intense simulated events would not be verified. We have found that no simulated HPEs occur among the False Alarms.

CHANGE: [13 282] "No HPE days belong to the misses...", add "and no simulated HPE days belong to the false alarms."

15 306ff As already mentioned, differences could results from the parameterization schemes as convection could not be resolved by the model. Also initial conditions like soil moisture have a significant influence (for references see main comment above)

RESPONSE: Thanks to the referee for giving some suggestions about the key factors associated with the positive S-component.

CHANGE: [15 308] "An hypothesis to explain such a result might be that in order to reach rainfall amounts that occurs in HPEs, the model needs to produce rainfall processes of larger extension."→ "Differences in A-component may result from the use of parameterizations, which leads to an underestimation of rainfall amounts. This deficiency may be related to the convection part not represented in the parametrization scheme. It may also be related to the representation of orography at a coarse resolution. As shown by Ehmele et al. (2015), an adequate representation of topographic features and local dynamic effects are required to correctly describe the interaction between orography and atmospheric processes. Furthermore, initial conditions have been shown to have a significant influence on rainfall forecasting (Kunz et al., 2018; Khodayar et al., 2018; Caldas-Álvarez et al., 2017)"

- Figure 7: Differences in A-component may result from the parameterization which lead to an underestimation of rainfall mounts. Deviations in the S-component can origin in misrepresentation of the orography and other local dynamic effects.

  RESPONSE: Thanks to the referee for providing physical explanations about the behaviour of S and A component.

  CHANGE: see previous suggestion of modification

- Table 4: The correlations are very weak and care has to be taken for the interpretation.

  RESPONSE: We agree that correlation is not very large. However, since the statistical test is significant, we could expect that these two quantities may be at least partially related

  CHANGE: [16 327] "Although correlations are statistically significant, it is worth noting that values are quite weak (in particular for cluster 5)."

16 325 'table 4': Table always with capital 'T'

  RESPONSE: Thanks to the referee for reporting this typo. The text has been corrected.

  CHANGE: As suggested by the reviewer

- Table 5: In general, this table is hard to read and understand. Which bracket belongs to which cluster? For scheme combinations that where used several times (e.g. B85) is it a mean value of all simulations? There are a very few cases with statistically significant differing distributions. It is also a bit confusing that one part of the table belongs to the A-component and the other part to the S-component. Same for Table 6. Maybe it is better to split this.

  RESPONSE: Thanks to the referee for this comment. In order to respond to the proposition of Referee 1 who asked to shorten the article, we have decided to remove these tables to make the article more legible. A list of the major modifications has been given in the first part of this document.

  CHANGE: Tables 5 and 6 are removed.

19 381ff Where can I find this? You say in Table 5 + 6, but it is not given which bracket belongs to which cluster. And how do I have to interpret the numbers to get this statement.

  RESPONSE: Thanks to the referee for this comment. Tables 5 and 6 are removed.

19 385ff Where can I find the numbers to prove this?

  RESPONSE: Thanks to the referee for this comment. Tables 5 and 6 are removed. The statement at line [19 385ff] is removed too.

20 400 'The departure from [...]', I think you mean 'The deviation from'

  RESPONSE: Thanks to the referee for this suggestion. The text has been corrected.

  CHANGE: As suggested by the reviewer

20 402ff Eq.(11)+(12) Are there other possibilities for the lower/upper boundary of the integral instead of -2 or +2? Where does this come from? Please specify.

RESPONSE: Thanks to the referee for requiring this specification. These boundaries are set since S and A components range by definition between these values. We noticed that a typo occurred. We used alternatively x or t in the integrals, whereas only one variable is required.

CHANGE: [20 400] "These functions are estimated over a bounded interval, corresponding to the finite range of S and A components." Eq. 11 and 12: t are replaced by x.

22 420ff '[...] the S-component exhibits the highest error on the right side of the distribution for B85 [...]', according to the given tables, this is not true for cluster 2 and LT34

RESPONSE: Thanks to the referee for noting this exception concerning the behaviour of the S-component. Since tables are replaced by the figure, this specific statement is modified and reformulated.

CHANGE: "In contrast to the A-component, the S-component exhibits the highest err_+ for B85 scheme for most of the cases (majority of + sign in Fig. (new figure)(b)), whereas this trend is not systematic for PCMT physics."

- Figure 11: Differences for dashed lines not visible. I would recommend a logarithmic y-axis or a separation into two y-axis (left and right)

  RESPONSE: Thanks to the referee for suggesting some modification to the plot. These plots are mainly conceived to highlight the differences in terms of absolute value between the first object (solid line) and the second object (dashed line). We believe that this difference should be less clear using a logarithmic y-axis or a second axis.

25 446 too many brackets in a row

RESPONSE: Thanks to the referee for reporting this typo.

CHANGE: Extra brackets have been removed

- Figure 12: I wonder what is about objects that are larger than the investigation area?

  RESPONSE: The large extension of the domain compared to the small-sized geographical features results in objects smaller than the total extension of the domain of interest for the majority of the dates over the period. However, it may happen that some objects that extend outside of the domain of concern are limited by the boundaries.

  CHANGE: [10 222] "Although objects are smaller than the domain for most of the situations, a few objects extending outside the domain are consequently limited by the boundaries of the region concerned."

28 480ff Following Fig. 13, there is an underestimation of the model compared to the observations for cluster 5 and a huge overestimation for cluster 2.

Only for cluster 3 the distributions look similar over the total range. So the statement given here is imprecise.

RESPONSE: Thanks to the referee for suggesting a more accurate specification. We agree that an overestimation concerns few extreme cases of cluster 2 and an underestimation is observed for cluster 5, characterising a very small portion of the distribution of observed pattern rainfall amounts. Except for these deviations, distributions seem to match each other.

CHANGE: [28 479] "For the most extreme clusters, object mass distribution of physics is similar to the distribution drawn from the observation, especially for cluster 5." → "Except for some deviation concerning a few extreme cases of cluster 2 and a small portion of distributions of cluster 5, object mass distribution of physics is similar to the distribution drawn from the observation, especially for cluster 3."

References

AghaKouchak, A., Behrangi, A., Sorooshian, S., Hsu, K., and Amitai, E. (2011). Evaluation of satellite-retrieved extreme precipitation rates across the central United States. *Journal of Geophysical Research: Atmospheres*, 116(D2).

Argence, S., Lambert, D., Richard, E., Chaboureau, J.-P., and Söhne, N. (2008). Impact of initial condition uncertainties on the predictability of heavy rainfall in the Mediterranean: A case study. *Quarterly Journal of the Royal Meteorological Society*, 134(636):1775–1788.

Boisserie, M., Descamps, L., and Arbogast, P. (2015). Calibrated Forecasts of Extreme Windstorms Using the Extreme Forecast Index (EFI) and Shift of Tails (SOT). *Weather and Forecasting*, 31(5):1573–1589.

Buizza, R. and Palmer, T. N. (1995). The Singular-Vector Structure of the Atmospheric Global Circulation. *Journal of the Atmospheric Sciences*, 52(9):1434–1456.

Caldas-Álvarez, A., Khodayar, S., and Bock, O. (2017). Gps – zenith total delay assimilation in different resolution simulations of a heavy precipitation event over southern france. *Advances in Science and Research*, 14:157–162.

Charron, M., Pellerin, G., Spacek, L., Houtekamer, P. L., Gagnon, N., Mitchell, H. L., and Michelin, L. (2009). Toward Random Sampling of Model Error in the Canadian Ensemble Prediction System. *Monthly Weather Review*, 138(5):1877–1901.

Collier, C. G. (2007). Flash flood forecasting: What are the limits of predictability? *Quarterly Journal of the Royal Meteorological Society*, 133(622):3–23.

Davis, C., Brown, B., and Bullock, R. (2006). Object-Based Verification of Precipitation Forecasts. Part I: Methodology and Application to Mesoscale Rain Areas. *Monthly Weather Review*, 134(7):1772–1784.

Davis, C. A., Brown, B. G., Bullock, R., and Halley-Gotway, J. (2009). The Method for Object-Based Diagnostic Evaluation (MODE) Applied to Numerical Forecasts from the 2005 NSSL/SPC Spring Program. *Weather and Forecasting*, 24(5):1252–1267.

Descamps, L., Labadie, C., and Bazile, E. (2011). Representing model uncertainty using the multiparametrization method. In *Workshop on Representing Model Uncertainty and Error in Numerical Weather and Climate Prediction Models, 20-24 June 2011*, pages 175–182, Shinfield Park, Reading. ECMWF, ECMWF.

Descamps, L., Labadie, C., Joly, A., Bazile, E., Arbogast, P., and Cébron, P. (2015). PEARP, the Météo-France short-range ensemble prediction system. *Quarterly Journal of the Royal Meteorological Society*, 141(690):1671–1685.

Du, J., Mullen, S. L., and Sanders, F. (1997). Short-Range Ensemble Forecasting of Quantitative Precipitation. *Monthly Weather Review*, 125(10):2427–2459.

Ebert, E. E. and McBride, J. L. (2000). Verification of precipitation in weather systems: Determination of systematic errors. *Journal of Hydrology*, 239(1):179–202.

Ehmele, F., Barthlott, C., and Corsmeier, U. (2015). The influence of sardinia on corsican rainfall in the western mediterranean sea: A numerical sensitivity study. *Atmospheric Research*, 153:451 – 464.

Hamill, T. M. (2012). Verification of TIGGE Multimodel and ECMWF Reforecast-Calibrated Probabilistic Precipitation Forecasts over the Contiguous United States. *Monthly Weather Review*, 140(7):2232–2252.

Hamill, T. M., Hagedorn, R., and Whitaker, J. S. (2008). Probabilistic Forecast Calibration Using ECMWF and GFS Ensemble Reforecasts. Part II: Precipitation. *Monthly Weather Review*, 136(7):2620–2632.

Hamill, T. M. and Whitaker, J. S. (2006). Probabilistic Quantitative Precipitation Forecasts Based on Reforecast Analogs: Theory and Application. *Monthly Weather Review*, 134(11):3209–3229.

Houtekamer, P. L., Lefaivre, L., Derome, J., Ritchie, H., and Mitchell, H. L. (1996). A System Simulation Approach to Ensemble Prediction. *Monthly Weather Review*, 124(6):1225–1242.

Houtekamer, P. L. and Mitchell, H. L. (1998). Data Assimilation Using an Ensemble Kalman Filter Technique. *Monthly Weather Review*, 126(3):796–811.

Khodayar, S., Czajka, B., Caldas-Alvarez, A., Helgert, S., Flamant, C., Di Girolamo, P., Bock, O., and Chazette, P. (2018). Multi-scale observations of atmospheric moisture variability in relation to heavy precipitating systems in the northwestern mediterranean during hymex iop12. *Quarterly Journal of the Royal Meteorological Society*, 144(717):2761–2780.

Kunz, M., Blahak, U., Handwerker, J., Schmidberger, M., Punge, H. J., Mohr, S., Fluck, E., and Bedka, K. M. (2018). The severe hailstorm in southwest germany on 28 july 2013: characteristics, impacts and meteorological conditions. *Quarterly Journal of the Royal Meteorological Society*, 144(710):231–250.

Lalaurette, F. (2003). Early detection of abnormal weather conditions using a probabilistic extreme forecast index. *Quarterly Journal of the Royal Meteorological Society: A journal of the atmospheric sciences, applied meteorology and physical oceanography*, 129(594):3037–3057.

Mass, C. F., Ovens, D., Westrick, K., and Colle, B. A. (2002). Does increasing horizontal resolution produce more skillful forecasts? *Bulletin of the American Meteorological Society*, 83(3):407–430.

Mittermaier, M., North, R., Semple, A., and Bullock, R. (2015). Feature-Based Diagnostic Evaluation of Global NWP Forecasts. *Monthly Weather Review*, 144(10):3871–3893.

Molteni, F., Buizza, R., Palmer, T. N., and Petroliagis, T. (1996). The ECMWF Ensemble Prediction System: Methodology and validation. *Quarterly Journal of the Royal Meteorological Society*, 122(529):73–119.

Nuissier, O., Ducrocq, V., Ricard, D., Lebeaupin, C., and Anquetin, S. (2008). A numerical study of three catastrophic precipitating events over southern France. I: Numerical framework and synoptic ingredients. *Quarterly Journal of the Royal Meteorological Society*, 134(630):111–130.

Palmer, T., Buizza, R., Doblas-Reyes, F., Jung, T., Leutbecher, M., Shutts, G., Steinheimer, M., and Weisheimer, A. (2009). Stochastic parametrization and model uncertainty. *ECMWF Technical Memorandum*, (598):42.

Petroliagis, T., Buizza, R., Lanzinger, A., and Palmer, T. N. (1997). Potential use of the ECMWF Ensemble Prediction System in cases of extreme weather events. *Meteorological Applications*, 4(1):69–84.

Ramos, A. M., Trigo, R. M., Liberato, M. L. R., and Tomé, R. (2015). Daily precipitation extreme events in the iberian peninsula and its association with atmospheric rivers. *Journal of Hydrometeorology*, 16(2):579–597.

Ricard, D., Ducrocq, V., and Auger, L. (2011). A Climatology of the Mesoscale Environment Associated with Heavily Precipitating Events over a Northwestern Mediterranean Area. *Journal of Applied Meteorology and Climatology*, 51(3):468–488.

Rossa, A., Nurmi, P., and Ebert, E. (2008). Overview of methods for the verification of quantitative precipitation forecasts. In Michaelides, S., editor, *Precipitation: Advances in Measurement, Estimation and Prediction*, pages 419–452. Springer Berlin Heidelberg, Berlin, Heidelberg.

Schumacher, R. S. and Davis, C. A. (2010). Ensemble-Based Forecast Uncertainty Analysis of Diverse Heavy Rainfall Events. *Weather and Forecasting*, 25(4):1103–1122.

Stensrud, D. J., Brooks, H. E., Du, J., Tracton, M. S., and Rogers, E. (1999). Using Ensembles for Short-Range Forecasting. *Monthly Weather Review*, 127(4):433–446.

Toth, Z. and Kalnay, E. (1993). Ensemble Forecasting at NMC: The Generation of Perturbations. *Bulletin of the American Meteorological Society*, 74(12):2317–2330.

Toth, Z. and Kalnay, E. (1997). Ensemble Forecasting at NCEP and the Breeding Method. *Monthly Weather Review*, 125(12):3297–3319.

Vié, B., Nuissier, O., and Ducrocq, V. (2010). Cloud-Resolving Ensemble Simulations of Mediterranean Heavy Precipitating Events: Uncertainty on Initial Conditions and Lateral Boundary Conditions. *Monthly Weather Review*, 139(2):403–423.

Walser, A., Lüthi, D., and Schär, C. (2004). Predictability of Precipitation in a Cloud-Resolving Model. *Monthly Weather Review*, 132(2):560–577.

Walser, A. and Schär, C. (2004). Convection-resolving precipitation forecasting and its predictability in Alpine river catchments. *Journal of Hydrology*, 288(1):57–73.

Wernli, H., Paulat, M., Hagen, M., and Frei, C. (2008). SAL—A Novel Quality Measure for the Verification of Quantitative Precipitation Forecasts. *Monthly Weather Review*, 136(11):4470–4487.

World Meteorological Organization, editor (2012). *Guidelines on Ensemble Prediction Systems and Forecasting*. Number 1091. WMO.