# Response to Reviewer 1

February 2020

## 1 General comment

The manuscript analyses the quality of precipitation reforecasts over the Mediterranean region performed with the Arpege model and a 10-member ensemble consisting of different convection parametrization schemes. The statistical data analysis is of very high technical quality. However the manuscript is very long, is sometimes a bit too technical, a large list of results from different well elaborated tools - one would wish to see a bit more physical interpretations in places - and one might possibly also see in addition to the observations some other reference, in particular the ERA5. Therefore I suggested major revisions, but as you will see these are not really major and should be very straightforward to do as it consist of some shortening as specified, a few sentences on physical interpretation and if possible I and likely the community would like to see on at least a few plots the results from the ERA5 (at least for forecast step 12 or so) which should shed more light on the quality/interpretation of the Arpege reforecasts (though they are somewhat penalized by initialising with a low-resolution ERA-Interim).

RESPONSE: Thanks to the referee for his general comment. First we agree with the fact that the manuscript could be improved and shortened. The revised manuscript has been shortened with more interpretation of the results.
Concerning ERA5, if the suggestion is to use ERA5 as a bench-marking dataset of our 30-year reforecast with its 10-member ensemble, it is not directly feasible in our sense. Indeed our study focuses on 24h precipitation fields whereas ERA5 10-member ensemble only provides 18h forecasts for which the observation density is too low. Our intention was not precisely to emphasize the relative quality of the PEARP system in comparison with other ensemble systems. It has been already done even though on much shorter dataset. It is more about to understand the inner behaviour of the PEARP error model representation in term of predictability in the case of HPEs.
Another technical limitation to undertake such a comparison comes, as stated by the referee, from the differences in resolution of the two systems, 10 kms for PEARP and around 25 kms for ERA5. Finally, the differences in the intialization, ERA-interim with a 75kms resolution and ERA5-25kms in the case of ERA5, could also be a sensitive factor of deviation in the comparison of the two systems.

## 2 Major comments

- add on a few plots at least the results with the ERA5 -remove Tables 7-9 or alternatively if you prefer Tables 5-6 -remove Figures 11 and 12 and corresponding text on page 25

  RESPONSE: The answer to ERA5 suggestion has been developed in the major comment answer (see previous section). We agreed with the reviewer suggestion to remove some of the tables which were too technical and the text has been simplified leading to an overall reduced length of the manuscript. Some too technical parts have also been removed.

  CHANGE: In the conclusion section (extensively modified, at the suggestion of referee number 2), a statement concerning ERA5 is added: "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."

  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.

  In order to simplify the text and facilitate the interpretation of results, Tables 7,8,9, and 10 have been replaced by a new figure attached to the document.

  We agree also with the reviewer that Fig. 12 and its corresponding comments were not necessary for the results analysis whole comprehension and hence they have been removed.

## 3 Minor comments

- Abstract l4 "flash-flood"→"flash floods" l6 "hindcast"→"hindcasts" -l27 :"a quasi-stationary synoptic system to slow the convective system", odd sentence, the convective systems are not "slowed" by the convective system

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

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

- l45 "weather warning triggering"→"triggering of weather warnings"

  RESPONSE: The text has be corrected

  CHANGE: As suggested by the reviewer

- -l86 "forecasts on the basis of the region of the domain ..." ????? rewrite whole sentence

  RESPONSE: Thanks to the referee for suggesting a reformulation of the sentence. The text has be corrected

  CHANGE: "In particular we focus on intense precipitation over the French Mediterranean region."

- -l91 "In detail, section 2.1 .."→ "section 2.1 .."

  RESPONSE: The text has be corrected

  CHANGE: as suggested by the reviewer

- -l95 ", and furtherly based on ..." delete

  RESPONSE: The text has be corrected

  CHANGE: As suggested by the reviewer

- "In ECUME oceanic fluxes are maximized"? what do you mean by that

  RESPONSE: thanks to the referee for requiring a clarification. A parameter that control evaporation fluxes is modified. The sentence is modified to clarify the discrepancy between $ECUME$ and $ECUME_{mod}$

  CHANGE: "In $\mathrm{ECUME_{mod}}$ evaporation fluxes above sea surfaces are enhanced

- l135 "realized"→"provided" -l232 "both in" delete -page 12 caption Figure 4 "the the" delete "the" -l312 "higher lead time ranges"→"longer lead times" -l349 "at first lead times"→"at short lead times"

  RESPONSE: Thanks to the referee for suggesting some corrections. The text has be corrected

  CHANGE: As suggested by the reviewer

- -l356 "while CAPE shows a behaviour different from". Can you explain why and add to the text?

  RESPONSE: Thanks to the referee for suggesting a specification about the interpretation of the behaviour of CAPE. The distribution of CAPE physics for L-component actually exhibits a different behaviour compared to the other ones. This effect may be related to the closure used in the deep convection parametrization scheme. CAPE implements the same scheme as B85, but it uses a closure based on CAPE. This discrepancy affects the spatial distribution of precipitation.

  CHANGE: "The use of a closure based on CAPE, rather than on the convergence of humidity seems to modulate the location of precipitation produced by this deep convection parametrization scheme. Moreover, at LT34 CAPE is characterized by a lower number of strong location errors, compared to the other physics."

- -l375 "standardized"→"normalized" -l503 "flat and large objects"→"flat and larger"

  RESPONSE: Thanks to the referee for suggesting some corrections. The text has be corrected

  CHANGE: As suggested by the reviewer

- -l504-5 please rewrite this sentence

  RESPONSE: Thanks to the referee for this comment. The sentence is not clear, indeed. The text has be corrected

  CHANGE: "As in grid-point rainfall verification, all the SAL components get worse as a function of lead time"

- 506 "negative-skewed"→"negatively-skewed"

  RESPONSE: The text has be corrected

  CHANGE: As suggested by the reviewer

- -l508 " .. component behaviour .. to the clusters" rewrite this sentence

  REPONSE: Thanks to the referee for this comment. The sentence is not clear, indeed. The text has be corrected

  CHANGE: "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."

- –l520-524 delete as with material Figures 11,12

  RESPONSE: Thanks to the referee for suggesting a shortening of this article. As already stated in response of the major comment, we deleted Fig. 12 and the corresponding text

  CHANGE: Remove Fig. 11. Remove text from [25 448] to [25 463]. Remove text [25 465]: "even when its scaled object mass $R^*$ is close to $O_2$"

- -l525 "On another hand"→"Furthermore," -l515 "ranges higher"→"ranges longer"

  RESPONSE: The text has be corrected

  CHANGE: As suggested by the reviewer

# References

# Response to Reviewer 2

February 2020

## 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.

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

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.

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.

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.

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.

# Systematic errors analysis of heavy precipitating ~~events~~ event prediction using a 30-year hindcast dataset

Matteo Ponzano[1], Bruno Joly[1], Laurent Descamps[1], and Philippe Arbogast[1]

[1]CNRM, Météo-France, Toulouse, France

**Correspondence:** Matteo Ponzano (matteo.ponzano@meteo.fr)

**Abstract.** The western Mediterranean region is prone to devastating ~~flash-flood~~ flash floods induced by heavy precipitation events (HPEs), which are responsible for considerable human and material damage. Quantitative precipitation forecasts have improved dramatically in recent years to produce realistic accumulated rainfall estimations. Nevertheless, ~~challenging issues remain in reducing~~ there are still challenging issues which must be resolved to reduce uncertainties in the initial conditions assimilation and the modeling of physical processes. In this study, ~~the spatial errors resulting from a 30-year (1981-2010) ensemble hindcast which implement the same physical parametrizations as in the operational~~ we analyze the HPE forecasting ability of the multi-physics based ensemble model operational at Météo-France ~~short-range ensemble prediction system,~~ Prévision d'Ensemble ARPEGE (PEARP)~~, are analysed. The hindcast consists of a 10-member ensemble reforecast, run every 4-days, covering the period from September to December.~~ . 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 ~~fields are classified~~ 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 ~~the HPEs. The~~ 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 ~~is then~~ that is performed on the model forecast and reference rainfall fields~~, which shows that both~~ . 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 ~~components~~ of the rainfall patterns are basically driven by the deep convection parametrization. Between the two main deep convection schemes used in PEARP, we qualify that the PCMT parametrization scheme performs better than the B85 scheme. A further analysis of spatial features of the rainfall objects to which the SAL metric pertains shows the predominance of large objects in the verification measure. It is for the most extreme events that the model has the best representation of the distribution of object integrated rain.

## 1 Introduction

Episodes of intense rainfall in the Mediterranean affect ~~western Europe climate~~ the climate of western Europe and can have important societal impact. During these events, daily rainfall amounts associated ~~to a one~~ with a single event can reach annual equivalent values. These rainfall events coupled with a steep orography are responsible for associated torrential floods, which may cause considerable human and material damage. In particular, Southern France is prone to devastating ~~flash flood~~ flash flood events such as the Aude case (Ducrocq et al., 2003), Gard (Delrieu et al., 2005), and Vaison-La-Romaine (Sénési et al., 1996), which occurred on 12–13 November 1999, 22 September 1992 and 8-9 September 2002, respectively. For instance, in the Gard case more than 600 mm were observed locally during a two-day event and 24 people were killed during the associated flash flooding. Extreme rainfall ~~amounts~~ events generally occur in a synoptic environment favourable for such events (Nuissier et al., 2011).

A detailed list of the main ~~basic atmospheric ingredients~~ atmospheric factors which contribute to the onset of HPEs ~~(Heavy Precipitation Event)~~ are reported by Lin et al. (2001): 1) a conditionally or potentially unstable airstream impinging on the mountains, 2) a very moist low-level jet, 3) a steep mountain, and 4) a quasi-stationary ~~synoptic system to slow the convective system~~ convective system that persists over the threat area. In ~~the~~ Southeastern France, the Mediterranean Sea acts as a source of energy and moisture ~~to the lower levels and a pronounced orography is present~~ which is fed to the atmospheric lower levels over a wide pronounced orography above the Massif Central, Pyrenees, and South Alps areas (Delrieu et al., 2005). Extreme rainfall amounts are enhanced especially along the Southern and Eastern foothills of mountainous chains (Frei and Schär, 1998; Nuissier et al., 2008), in particular the Southeastern part of the Massif Central (Cévennes). ~~Nevertheless, all the~~ 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 ~~in order to emphasize the climatological mesoscale environment~~ and showed the existence of mesoscale environments associated with heavy precipitating events. ~~They considered~~ Considering four sub-domains~~according to the location of precipitation. They~~ , they found that the synoptic and mesoscale patterns can greatly differ as a function of the location of the precipitation.

~~A HPE could be convective (or not) or a combination of both (Ducrocq et al., 2002). Extreme rainfall amounts~~ Extreme rainfall events are generally associated with coherent structures slowed down and enhanced by the relief, whose extension ~~must be~~ is often larger than a single thunderstorm cell. ~~Mesoscale processes can be crucial in organizing a very large variety of precipitating systems. For some cases, a~~ At some point, this mesoscale organization can turn into a self-organization process leading to a mesoscale convective system (MCS) ~~can stay stationary for several hours, affecting a limited area. Quasi-stationary MCSs are particularly efficient in terms of rain production due to their high intensities and their spatial stationarity (Nuissier et al., 2008). This stationarity is explained by the regeneration of new convective cells at a rate compensating the advective speed of the older cells (Ducrocq et al., 2008).~~ when interacting with their environment, which in turn leads to high intensity rainfall (Nuissier et al., 2008).

If long term territory 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 anticipating earlier alert would be beneficial in a the short term. Weather forecasting coupled with hydrological impact forecast forecasting is the main source of information for weather warning triggering triggering of weather warnings. Severe weather warnings are issued for the 24-hours 24-hour forecast only. However, in some caseseven at 24 hours term, the forecast process could be based on a model analysis issued some days prior to the issuance of the severe weather warnings. A better understanding of the sources of model uncertainty at such time-range may represent provide a major source of improvement for early diagnosis.

Uncertainties of initial conditions and Forecast uncertainties can be related to initialization data (analysis) or lateral boundary conditionsfor HPEs can be investigated for , and it has been investigated with both deterministic models (Argence et al., 2008) or and ensemble models (Vié et al., 2010). Several journal articles studied the predictability associated to showed that predictability associated with intense rainfall and flash floods (Walser et al., 2004; Walser and Schär, 2004; Collier, 2007). They showed that predictability limitations increase rapidly with decreasing scale since individual convective cells are rendered unpredictable by chaotic aspects of the moist dynamics. Moreover Quantitative Precipitation Forecast (QPF) appears to be more predictable in mountainous areas, where the triggering of convection and the larger-scale uplift results in a topographic control of the precipitation. Probabilistic forecast,

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 , is a suitable tool to explore the source of uncertainty for the predictability of HPEs (Du et al., 1997; Petroliagis et al., 1997; Stensrud et al., 1999; Schumacher and Davis, 2010; World Meteorological Organization, 2012) . An ensemble forecast consists of several realizations of the evolution of the state of the atmosphere, in order to assess

have shown the general ability of such models to sample 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 to the forecast . Forecast uncertainty is a mix of with the forecast is usually assessed by taking into account initial and model errors propagation. Major meteorological centres implemented different methods in order to take into account initial errors, 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). Model errors can be simulated through a multimodel approach, adding a stochastic component to the tendencies from parametrization schemes (Palmer et al., 2009), stochastically backscattering energy into the model (Berner et al., 2009)

~~or using~~ 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 ~~for~~ in each forecast member

~~(multiphysics approach: Charron et al., 2009; Descamps et al., 2011)~~

(Charron et al., 2009; Descamps et al., 2011).

The ~~framework set-up implemented for this study is a reforecast dataset built from a simplified version of the operational~~ global ensemble model implemented at Météo-France ~~short-range ensemble prediction system,~~ Prévision d'Ensemble ARPEGE (PEARP; Descamps et al., 2015) ~~, in which only model uncertainties are represented, by means of~~ is based on the second technique, also known as a multi-physics approach.

~~A reforecast ensemble dataset can be used for the calibration of the related version of the operational model (Hamill and Whitaker, 2006; Hamill et al., 2008; Hamill, 2012; Boisserie et al., 2015). Reforecast datasets have also been used to perform characterisation of a parameter forecast extremeness relatively to a reference, by comparing ensemble distributions to reforecast distribution like in Extreme Forecast Index computations (Boisserie et al., 2015; Lalaurette, 2003) . In this study, the production of a 30-year reforecast dataset provides a statistical basis for the exploration of the climatology~~

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~~configurations implemented in the operational ensemble system. Moreover this large dataset, spreading out over a multidecadal period, may include a significant number of intense events. We adopt a 10-km grid spacing reforecast ensemble to emphasize the predictability of mesoscale events rather than scattered and isolated phenomena, which are better represented by high-resolution models. The use of a coarser model resolution ensures a longer time integration for a given computing power. Consequently, predictability can be investigated up to 4 days lead time~~. 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.

~~Traditional~~ In order to achieve such a systematic analysis, standard rainfall verification methods can be ~~exploited in order to assess the quality of a forecast as they are generally built on the basis of a~~ used. They are usually based on grid-point based ~~approach~~approaches. These techniques, especially when applied to intense events, are subject to ~~timing~~ time or position errors leading to low scores (Mass et al., 2002) ~~. This combination of both spatial and timing errors is~~ also known as the double penalty problem (Rossa et al., 2008). ~~Spatial~~ To counteract this problem, spatial verification techniques have been developed with the goal ~~to evaluate forecast quality in a manner similar to a forecaster approach and to overcome the traditional grid-point to grid-point verification limitations. A branch of spatial techniques is represented by the~~ 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., 2006a, 2009; Mittermaier et al., 2015; Wernli et al., 2008).~~In this study, the~~ The feature-based quality measure SAL ~~(Wernli et al., 2008, 2009) is used.~~ (Wernli et al., 2008, 2009) is 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.

~~The aim of this paper is to suggest a~~ 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~~ using an object oriented approach. In particular we focus on intense precipitation over the French Mediterranean region. In addition to the ~~quality of the spatial forecasts on the basis of the regionof the domain affected by the precipitations. Besides the~~ analysis of diagnostics from the SAL-metric, a statistical analysis of ~~the~~ 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. ~~In detail, section~~ Section 2.1 describes the reforecast ensemble dataset and section 2.2 ~~, the generation~~ details the creation of the daily rainfall reference~~and the statistical stratification of this product by means of a peak-over-threshold method and a~~ , 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 ~~separated SAL diagnostics for each physical parameterization scheme~~ SAL diagnostics divided into all different physical parametrization schemes of the ensemble reforecast, and ~~furtherly based on individual objects spatial properties~~ for the spatial properties of individual objects. Conclusions are given in section 4.

## 2 Data and methodology

### 2.1 PEARP hindcast

The PEARP reforecast dataset consists of a 10-member ensemble computed daily from 1800 UTC initial conditions, covering four ~~month~~ months (from September to December), every year of a 30-year period (1981-2010). This period has been chosen since ~~HPEs occurence~~ HPE occurrence in the region considered is largest during the autumn season (see Fig. 3 from Ricard et al., 2011). It uses ARPEGE (Action de Recherche Petite Echelle Grande Echelle, Courtier et al. (1991)), the global operational model of Météo-France with a spectral truncation T798, 90 levels on the vertical, and a variable horizontal resolution (mapping factor of 2.4 with a highest resolution of 10 km over France). One ensemble forecast is performed every 4 days of the four-month

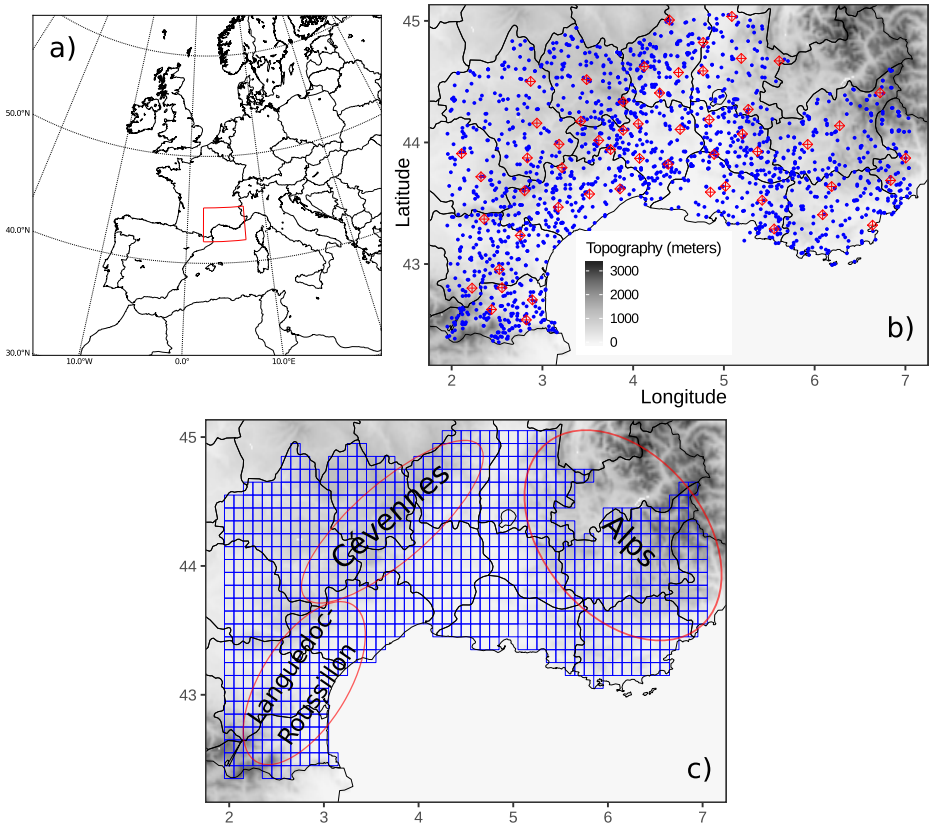


**Figure 1.** Panel **a** shows a situation map of the investigated area (rectangle with red edges) with respect to Western Europe and the Mediterranean Sea. Panel **b** shows the rain-gauges network used for the study. Red diamonds represent the rain-gauges selected for cross-validation testing, blue dots represent the rain-gauges selected for cross-validation training. Panel **c** shows the $0.1° \times 0.1°$ model grid (in blue), ~~corresponding to the domain D,~~ along with the location of three key areas. The domain D is located within the borders of the model grid (panel **c**).

period up to 108-hour lead time. Our initialization strategy follows the hybrid approach described in (Boisserie et al., 2016), in which first the atmospheric initial conditions are extracted from the ERA-Interim reanalysis (Dee et al., 2011) available at the European Center for Medium-range Weather Forecasts. Second, the land-surface initialization parameters are interpolated from an offline simulation of the land-surface SURFEX model (Masson et al., 2013) driven by the 3-hourly near-surface atmospheric fields from ERA-Interim. 24-hour accumulated precipitation forecasts are extracted on a $0.1° \times 0.1°$ grid, that defines the domain D (see Fig. 1c), which encompasses Southeastern France (Fig. 1a). The reforecast dataset does not have any representation of initial uncertainty, but it implements the same representation of model uncertainties (multiphysics approach) as in the PEARP operational version of 2016.

**Table 1.** Physical parametrizations used in the ensemble reforecast.

| | Turbulence | Shallow convection | Deep convection | Oceanic flux |
|---|---|---|---|---|
| Ref | TKE | KFB | B85 | ECUME |
| 1 | TKE | KFB | B85 | $ECUME_{mod}$ |
| 2 | L79 | KFB | $B85_{mod}$ | ECUME |
| 3 | L79 | KFB | CAPE | ECUME |
| 4 | $TKE_{mod}$ | KFB | B85 | ECUME |
| 5 | TKE | EDKF | B85 | ECUME |
| 6 | TKE | PMMC | PCMT | ECUME |
| 7 | TKE | KFB | PCMT | ECUME |
| 8 | TKE | PCMT | PCMT | ECUME |
| 9 | TKE | KFB | B85 | ECUME |

~~Nine~~ The same nine different physical parametrizations as the ones used in PEARP (see Table ~~1~~1) are added to the one corresponding to the ARPEGE deterministic physical package. Two turbulent diffusion schemes are considered: the Turbulent Kinetic Energy scheme (TKE; Cuxart et al., 2000; Bazile et al., 2012) and the Louis scheme (L79; Louis, 1979). $TKE_{mod}$ is a slightly modified version of TKE, in which horizontal advection is ignored. For shallow convection different schemes are used: a mass flux scheme introduced by Kain and Fritsch (1993) and modified by Bechtold et al. (2001), thereafter the KFB approach, the Prognostic Condensates Microphysics and Transport scheme (PCMT; Piriou et al., 2007)), the Eddy-Diffusivity/Kain-Fritsch scheme (EDKF) and the PMMC scheme (Pergaud et al., 2009). The deep convection component is parametrized by either the PCMT scheme or the Bougeault (1985) scheme (thereafter B85). Closing the equation system used in these two schemes means relating the bulk mass flux to the in-cloud vertical velocity through a quantity $\gamma$ qualifying the ~~area coverage of convection~~convection area coverage. Two closures are considered~~,~~: the first one (C1) is based on the convergence of humidity and the second one (C2) is based on the CAPE (Convective Available Potential Energy). B85 scheme originally uses the C1 closure, while PCMT ~~uses alternatively~~ alternatively uses the closure (C1 or C2) which maximizes the $\gamma$ parameter. Physics package 2 uses a modified version of the B85 scheme in which deep convection is triggered only if cloud top exceeds 3000 m ($B85_{mod}$ in Table 1). The same trigger is used in physics package 3 in which deep convection is parametrized using the B85 scheme along with a CAPE closure (CAPE in Table 1). Finally the oceanic flux is solved by means of the ECUME scheme (Belamari, 2005). In $ECUME_{mod}$ ~~oceanic fluxes are maximized~~evaporation fluxes above sea surfaces are enhanced. Control member and member 9 are characterized by the same parametrization set-up, but member 9 differs for the modelization of orographic waves.

### 2.2 Daily Rainfall Reference

24-hour accumulated precipitation is derived from the in-situ Météo-France rain-gauge network, covering the same period as the reforecast dataset. 24-hour rainfall amounts collected from fourteen French departments within the reforecast domain D are used (Fig. 1b). In order to maximize the rain-gauge network density within the region, all daily available validated data covering the period have been used.

Rain-gauge observations are used to build gridded precipitation references by a statistical spatial interpolation of the observations. The aim of this procedure is to ensure a spatial and temporal homogeneity of the reference, as well as the same spatial resolution as the reforecast dataset. Ly et al. (2013) ~~realized~~ provided a review of the different methods for spatial interpolation of rainfall data. They showed that kriging methods outperform deterministic methods for the computation of daily precipitation. However, both types of methods were found to be comparable in terms of hydrological ~~modeling~~ modelling results. For the interpolation, we use a mixed geo-statistical and deterministic algorithm, which implements Ordinary Kriging (OK; Goovaerts et al., 1997) and Inverse Distance Weighting methods (IDW, Shepard (1968)). For the kriging method three semi-variogram models (Exponential, Gaussian and Spherical) are fitted to daily sample semi-variogram drawn from raw and square root transformed data (G. Gregoire et al., 2008; Erdin et al., 2012). This configuration ~~implies~~ involves the use of six different geo-statistical interpolation models. In addition, four different IDW versions are used, by varying the geometric form parameter D used for the estimation of the weights (see eq. (2) in Ly et al. (2011)) and the maximum number $n$ of neighbour stations involved in the IDW computation. Three versions are defined by fixing $d = 2$ and alternatively assigning $n$ values equal to 5, 10 and $N$ (with $N$ being the total number of stations available for that specific day). In the fourth version we set $n = N$ and $d = 3$. For each day, a different interpolation method is used and its selection is based on the application of a cross validation approach. We select 55 rain-gauges as a training dataset (see the red diamonds in Fig. 1c) in order to have ~~a~~ sufficient coverage over the domain, especially on the mountainous area. Root Mean Square Error (RMSE) is used as a criterion of evaluation. For each day, the method which minimizes the RMSE computed within the rain-gauges of the training dataset is selected and the spatial interpolation is then performed on a regular high resolution grid of 0.05°. The highest resolution estimated points are then up-scaled to the 0.1°grid resolution of domain D, by means of a spatial average. This up-scaling procedure aims at reproducing the filtering effect produced by the parametrizations of the model on the physical processes that occur below the grid resolution.

#### 2.2.1 HPE database

We implement a methodology in order to select the HPEs from the daily rainfall reference. Anagnostopoulou and Tolika (2012) have examined parametric and non-parametric approaches for the selection of rare events sampled from a dataset. Here we adopt a non-parametric peak-over-threshold approach, on the basis of ~~the~~ WMO guidelines (World Meteorological Organization, 2016). The aim is to generate a set of events representative of the tail of the rainfall distribution for a given region and season. Following the recommendation of Schär et al. (2016), an all-day percentile ($P_{0 \leq n \leq 1}$) formulation is applied. A potential weakness of the research methodology based on the gridded observation reference is that a few extreme precipitation events affecting a

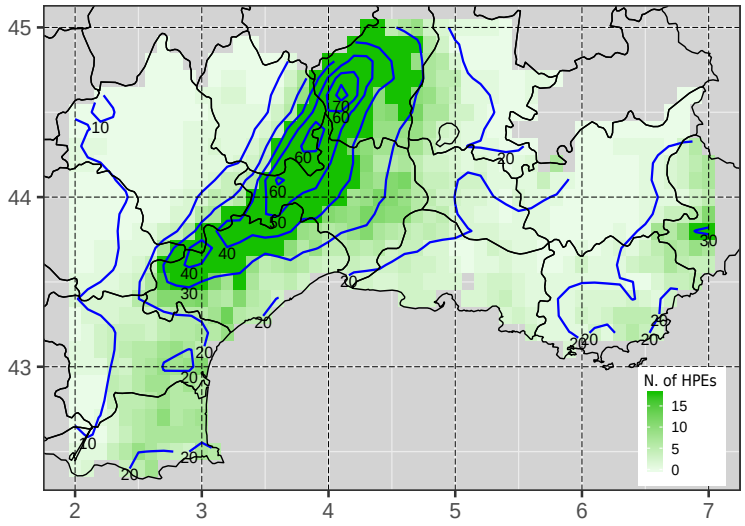


**Figure 2.** Annual average of ~~HPEs~~ HPE occurrence per grid point (in green). The composite of daily rainfall amounts (mm/day) of the HPE dataset is represented by the blue isohyets.

smaller area than the grid resolution may not be identified. However, this approach has been preferred to a classification using rain-rauges because spatial and temporal homogeneity are ensured.

We proceed as ~~follow:~~ follows: first the domain is split into two sub-regions based on the occurrence of climatological intense precipitations during the 30 ~~years~~ 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. ~~Then,~~ 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 ~~a HPE if , for that day, there exists~~ an HPE if one point of sub-region A ~~whose~~ accumulated rainfall is greater than 100 mm or if ~~there exists~~ one point of sub-region B ~~whose~~ 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. ~~Figure 2~~ 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 ~~is worth mentioning~~ should be noted that some points are never taken into account for the HPE selection (~~grey~~ white points of Fig. 2),

**Table 2.** Classification of days computed from 24-hour rainfall amounts in southern France (1981-2010), percentage of HPEs and fraction of HPEs. 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.

| **Cluster** | Total (%) | HPEs (%) | Fraction of HPEs (%) |
|---|---|---|---|
| 1 | 14.5 | 11.4 | 4.3 |
| 2 | 5.3 | 24.0 | 24.6 |
| 3 | 1.8 | 30.7 | 92.2 |
| 4 | 75.8 | 2.6 | 0.2 |
| 5 | 2.6 | 31.3 | 65.2 |
| *Total number of days* | 3660 | 192 | |

because the required conditions ~~are never satisfied.~~ 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).

### 2.2.2 Clustering analysis

Clustering analysis methods can be applied to daily rainfall amounts in order to identify emergent regional rainfall patterns. This classification is largely used for assessing the between-day spatial classification of heavy rainfall (Romero et al., 1999; Peñarrocha et al., 2002; Little et al., 2008; Kai et al., 2011). We applied a cluster analysis, as an exploratory data analysis tool, in order to assess geographical properties of the precipitation reference dataset. The size of the dataset is first reduced and the signal is filtered out by means of a principal component analysis (Morin et al., 1979; Mills, 1995; Teo et al., 2011). The first 13 Principal Components (PCs), whose projection explains $90\%$ of the variance, are retained. Then the $K$-means clustering method is applied. It is a non-hierarchical method based on the minimization of the intraclass variance and the maximization of the variance between each cluster. A characteristic of $k$-means method is that the number of clusters ($K$) into which the data will be grouped has to be *a priori* prescribed. Consequently, we ~~have first~~ first have to implement a methodology to find the number of clusters which leads to the most classifiable subsets.

The analysis is applied to the full reference dataset, including rainy and dry days. We run 2000 tests for a range of *a priori* cluster numbers $K$ that lie between 3 and 13, by varying a random initial guess each time. Then, for a given $K$, an evaluation of the stability of the assignment into each cluster is performed. The number of clusters is considered stable if each cluster size is almost constant from one test to another. $K = 5$ is retained as the most stable number of clusters and because it suggests a coherent regional stratification of the daily rainfall data. The final classification within the 2000 tests is selected by minimizing the sum of the distance between the cluster centroids from each test and the geometric medians of cluster centroids computed from all the tests. The test which minimizes this quantity has been selected as the reference classification. The results from the cluster classification are summarized in Table 2. The clusterization shows large differences in term of cluster size, more than

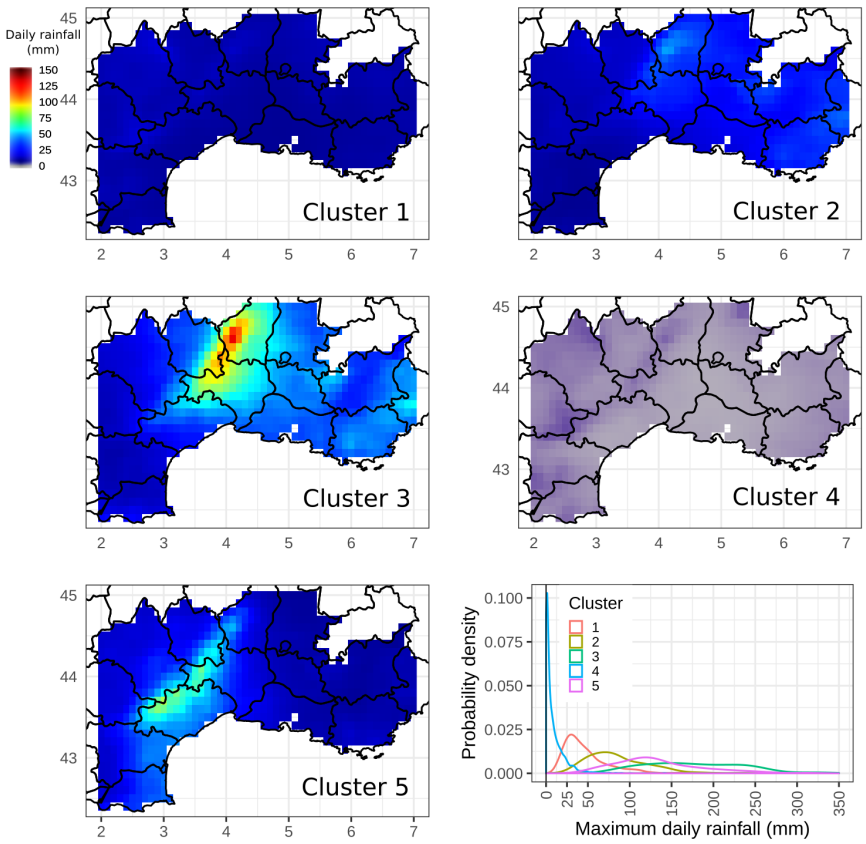


**Figure 3.** Rainfall composites (mm/day) for the 5 clusters selected by the *K*-means algorithm. The bottom-right panel shows the probability density distribution of the maximum daily rainfall (mm) for each cluster class.

3/4 of the dataset is grouped in cluster 4, which ~~collects mostly~~ mostly collects the days characterized by weak precipitation amounts or dry days. The percentage of HPEs within the clusters shows that the most intense events are represented in clusters 2, 3 and 5, among which cluster 5 ~~is the one with the~~ shows largest proportion of HPE (~~86~~65% of HPEs ~~days~~ within this cluster). Clusters 2,3 and 5 together account for 86% of the HPEs.

The same composite analysis as the one previously applied to HPE class, is now computed for each cluster class (Fig. 3). It ~~reveals~~ shows significant differences between clusters. Not only the relative intensity of events is different for each of the clusters, but also the location differs. Rainfall range is weak for cluster 1 and close to zero for cluster 4. Cluster 2 includes some moderate 24-hour rainfall amounts related to generalized precipitation events and a few ~~of~~ HPEs. For cluster 1, composite values are slightly higher on the northwestern area of the domain, while for cluster 2, rainfall amounts values are more ~~enhanced~~ significant on the eastern side of the domain D. Clusters 3 and 5 ~~contains together~~ together account for 63% of the HPEs of the whole period, but rainfall events seem to affect different areas. Cluster 3 includes most of the events impacting the Cévennes

mountains and the eastern departments on the southern side of the Alps. Cluster 5 average rainfall is enhanced along the southern side of the Cévennes, especially the Languedoc-Roussillon region.

The bottom-right panel of Fig. 3 shows the density distributions computed from the maximum daily rainfall for each cluster. It is worth noting ~~how each distribution samples different ranges of maximum grid-point~~ that cluster rainfall distributions cover different intervals of maximum daily rainfall amounts. Cluster 4 includes all the dry days. As this paper focuses on the most severe precipitation events, results will only be shown for ~~cluster~~ clusters 2, 3 and 5 for the remainder of the paper.

### 2.3 The SAL verification score

#### 2.3.1 The SAL score definition

The SAL score is an object-based quality measure introduced by Wernli et al. (2008) for the spatial verification of numerical weather prediction (NWP). It consists in computing three different components: structure **S** is a measure of volume and shape of the precipitations patterns, amplitude **A** is the normalized difference of the domain-averaged precipitation fields, and location **L** is the spatial displacements of patterns on the forecast/observation domains.

Different criteria for the identification of the precipitation objects could be implemented: a threshold level (Wernli et al., 2008, 2009), a convolution threshold (Davis et al., 2006a, b), or a threshold level conditioned to a cohesive minimum number of contiguous connected points (Nachamkin, 2009; Lack et al., 2010). The threshold level approach needs only one estimation parameter, so it has been preferred to the other methods for its simplicity and interpretability. Since we focus on the patterns associated ~~to~~ with the HPEs, we decided to adapt the threshold definition given by $T_f = x_{max} \times f$, where $x_{max}$ is the maximum precipitation value of the points belonging to the domain and $f$ is a constant factor ($= 1/15$, in the paper of Wernli et al. (2008)). Here the coefficient $f$ has been raised ~~up~~ to $1/4$, because a smaller value results in excessively large objects spreading out over most of the domain D. Choosing ~~an~~ a higher $f$ factor enables to obtain more realistic features within the ~~considered domain. Thresholds~~ domain considered. Threshold levels $T_f$ are computed daily for the reforecast and the reference dataset. 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.

If we consider the domain D, the amplitude A is computed as follows:

$$A = \frac{\langle R_{\text{for}} \rangle_D - \langle R_{\text{obs}} \rangle_D}{0.5\left(\langle R_{\text{for}} \rangle_D + \langle R_{\text{obs}} \rangle_D\right)} \in [-2, 2], \tag{1}$$

where $\langle \rangle_D$ denotes the average over the domain D. $R_{\text{for}}$ and $R_{\text{obs}}$ are the 24-hour rainfall amounts over D associated ~~to~~ with the forecast and the observation, respectively. A perfect score is achieved for $A = 0$. The domain-averaged rainfall field is overestimated by a factor 3 if $A = 1$, similarly it is underestimated by a factor 3 if $A = -1$. The amplitude is maximal ($A = 2$) if $\frac{\langle R_{\text{for}} \rangle_D}{\langle R_{\text{obs}} \rangle_D} \to +\infty$ and minimal ($A = -2$) if $\frac{\langle R_{\text{for}} \rangle_D}{\langle R_{\text{obs}} \rangle_D} \to 0$.

The two other components require the definition of precipitation objects (thereafter $\{Obj\}$), also called features, which represent contiguous grid points belonging to the domain D, characterized by rainfall values exceeding a given threshold. The location $L$ is a combined score defined by the sum of two contributions, $L1$ and $L2$. $L1$ measures the magnitude of the shift

between the center of mass of the whole precipitation field for ~~both in~~ the forecast ($\bar{x}_{\text{for}}$) and observation ($\bar{x}_{\text{obs}}$):

$$L1 = \frac{|\bar{x}_{\text{for}} - \bar{x}_{\text{obs}}|}{d} \in [0,1]\,, \tag{2}$$

where $d$ is the largest distance between two boundary points of the considered domain $D$. The second metric $L2$ takes into account the spatial distribution of the features inside the domain, that is the scattering of the objects:

$$r = \frac{\sum_{n=1}^{N} M_n |\bar{x} - x_n|}{\sum_{n=1}^{N} M_n}, \tag{3}$$

where $M_n$ is the integrated mass of the object $n$, $x_n$ is the center of mass of the object $n$, $N$ is the number of objects and $\bar{x}$ is the center of mass of the whole field.

$$L2 = 2\frac{|r_{\text{for}} - r_{\text{obs}}|}{d} \in [0,1]\,, \tag{4}$$

$$L = L1 + L2 \in [0,2]\,. \tag{5}$$

$L2$ aims at depicting ~~objects~~ object differences between observed and forecasted scattering of the precipitation objects. We can notice that the scattering variable (eq. (3)) is computed as the weighted distance between the center of total mass and the center of mass of each object. Therefore $L$ is a combination of the information provided by the global spatial distribution of the fields ($L1$) and the difference in ~~the~~ scattering of the features over the domain ($L2$). The location score is perfect if $L1 = L2 = 0$, so if $L = 0$ all the centers of mass match each ~~others~~other.

The S-component is based on the computation of the integrated mass $M_k$ of one object $k$, scaled by the maximum rainfall amount of the object $k$:

$$V_k = \frac{M_k}{\max\limits_{x \in Obj_k} R(x)} \frac{M_k}{\max R(x; x \in Obj_k)}. \tag{6}$$

Then, the weighted average $V$ of all features is computed, in order to obtain a scaled, weighted total mass:

$$V = \frac{\sum_{n=1}^{N} M_n V_n}{\sum_{n=1}^{N} M_n}, \tag{7}$$

$$S = \frac{V_{\text{for}} - V_{\text{obs}}}{0.5(V_{\text{for}} + V_{\text{obs}})} \in [-2,2]\,. \tag{8}$$

Then, $S$ represents the difference of both forecasted and observed volumes, scaled by their half-sum. It is important to scale the volume so that the structure is less sensitive to the mass, meaning that it relates more to the shape and extension of the features rather than their intensities. In particular $S < 0$ means that the forecast objects are large and/or flat compared to the observations. Inversely, peaked and/or smaller objects in the forecast give positive values of $S$. We refer to Wernli et al. (2008) for the exploration of the behaviour of SAL for some idealized examples.

On the basis of the definition of the score, it can be noticed that $A$ and $L1$ components are not affected by the object identification and depend only on the total rainfall fields.

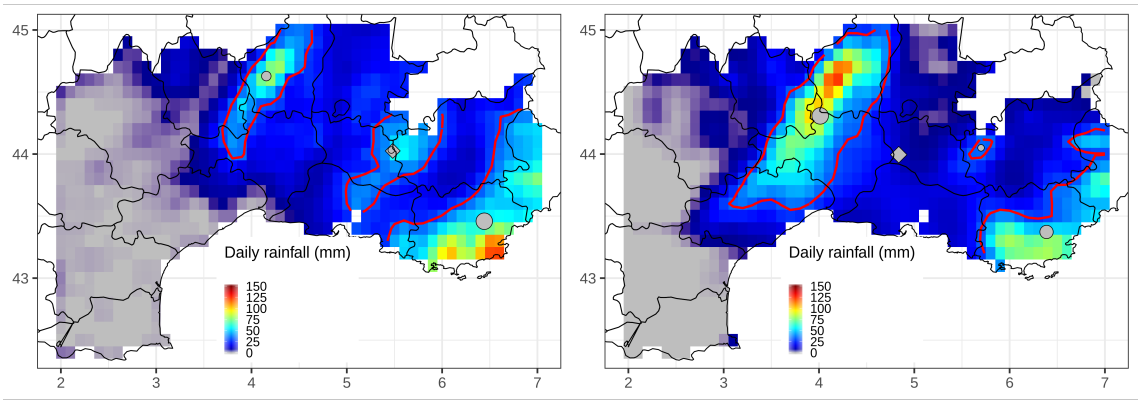


**Figure 4.** SAL pattern analysis for the case of 28 ~~october~~ October 2004, applied on the observation data (left panel), and one 60-hour lead time forecast (right panel). Base contour of the identified objects are in red lines. Gray points stand for the rain ~~barycentre~~ barycenter of each pattern, gray diamond depicts the rain barycenter for the whole field. ~~Size~~ The size of the ~~barycentre~~ barycenter points is proportional to the ~~the~~ integrated mass of the associated object.

### 2.3.2 A selected example of the application of SAL

An example of the SAL score applied to an HPE, that occurred on the 28 Oct 2004, is shown in Fig. 4 (60-hour lead time forecast run using the physical package n.8). For the rainfall reference, a 24-hour rainfall maximum value (121.3 mm), was registered in the ~~south-Eastern~~ southeastern coastal region. Therefore the threshold level $T_f$ is set to 30.3 mm. For the forecast, the maximum value is 123.1 mm ($T_f = 30.8$ mm) and, in contrast with the reference, it is located on the Cévennes. The number of objects, three, is equivalent in both fields. The value of $A$ is 0.08, which means that the domain-averaged precipitation field of the forecast is nearly similar to the reference one. The structure S-components is positive (0.28), which could be explained by the larger forecast object over the Cévennes area, while the object along the ~~south-eastern~~ southeastern coast is smaller and less intense. The contribution of the third object is negligible for the computation of $S$. The ~~location~~ L-component ~~$L$~~ is equal to 0.23, with $L1$=0.13 and $L2$=0.10. The location error $L1$ means that the distance between the ~~centres~~ centers of total mass (see diamonds in Fig. 4) is $13/100$ of the largest distance between two boundary points of the considered domain. This error is mostly due to the fact that the most intense rainfall patterns are far apart from each other in the observations and the forecast.

## 3 Analysis of the reforecast ~~HPEs~~ HPE representation

An SAL verification score has been applied to the reforecast dataset to perform statistical analysis of QPF errors. The reforecast dataset is considered as a testbed model in order to study sources of systematic errors ~~of~~ in the forecast. The overall reforecast performance is first examined for HPE/non-HPE~~cases~~, then according to the clusters. In a second step, the behaviour of the

**Table 3.** Contingency table computed for rainy and dry days.

| **Contingency table** | Obs rainy day | Obs dry day |
|---|---|---|
| Model rainy day | 3258 | 84 |
| Model dry day | 226 | 62 |

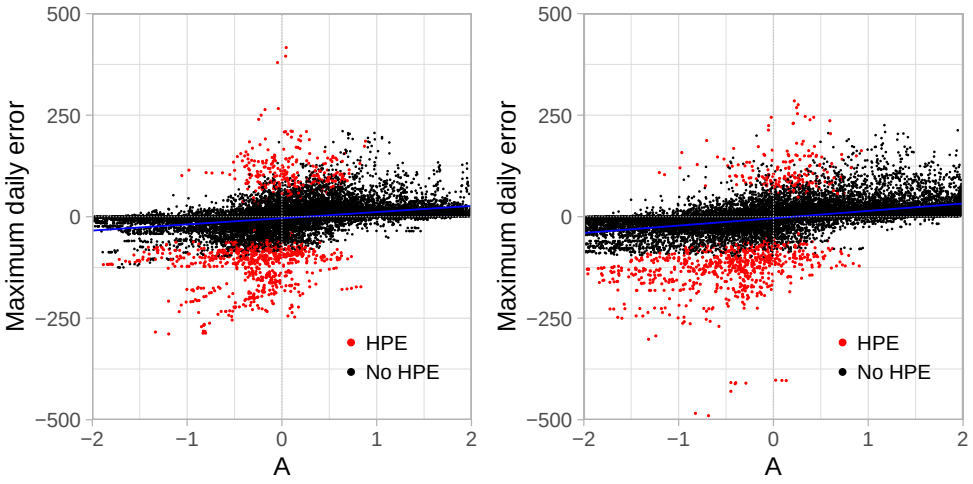


**Figure 5.** Relationship between the daily rainfall gridpoint maximum algebric error and the A-component of the SAL score. HPEs days are plotted in red, while other days are in black. Left panel is for LT12 lead time, right panel shows LT34 lead time. Linear regression analysis is added to the plot.

different physics schemes is ~~analysed by considering separately~~ analyzed by separately considering the SAL results of each reforecast member. Similarly, the analysis is again allocated to HPE/non-HPEs ~~cases~~ and subsequently to each cluster.

For both the reforecast and the reference, we set all the days with at least one grid point beyond 0.1 mm as a rainy day. In order to facilitate the comparison between the parametrizations, SAL verification is only performed when all the members and the reference are classified as rainy day. Table 3 shows the contingency table of the rainy and dry days. Therefore 84 false alarms, 226 missed cases, and 62 correctly forecast dry days are not involved in the SAL analysis. No HPE ~~days~~ belong to the misses and no simulated HPE belong to the false alarms. The SAL measure is then applied to the 3258 rainy days.

### 3.1 SAL Evaluation of the ~~HPEs~~ HPE forecast

#### 3.1.1 HPE/non-HPE~~cases~~

First the relationship between the A-component of SAL and the maximum grid-point error is investigated (Fig. 5). 36-hour and 60-hour lead times (LT12 hereafter) and 84-hour and 108-hour lead times (LT34 hereafter) are grouped together. Maximum daily absolute errors ~~ranges~~ range between -250 mm and 250 mm. Rare higher values are observed, which are likely related to strong

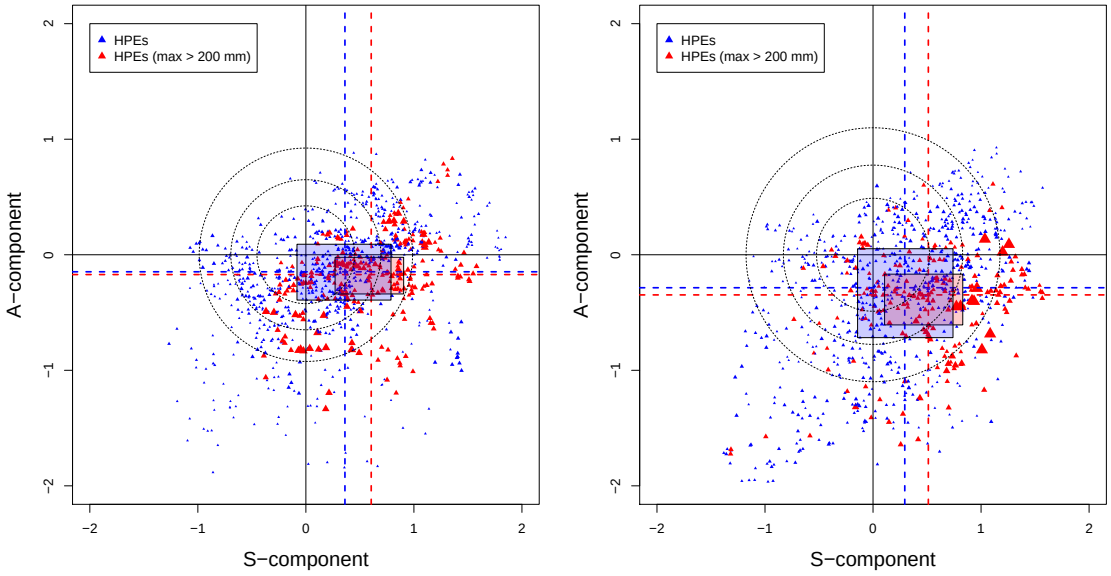


**Figure 6.** Relationship between the A-component and the S-component of the SAL score (SAL diagrams) for HPEs ~~events~~ only, for lead times LT12 (left) and LT34 (right). Blue triangles represent HPEs ~~events~~ with ~~rainfall~~ gridpoint maximum rainfall under 200 mm/day, and red triangles for rainfall ~~amount~~ amounts beyond 200 mm~~, triangles~~. Triangles are proportional to the rainfall value. Some main characteristics of the component distribution are plotted, the median value (dashed lines), percentile 25% and 75% delimitate the boxes. Circles represent the limits 25%, 50% and 75% percentiles to the best score (A=0, S=0).

double penalty effects that often occur in gridpoint-to-gridpoint verification. Points are mostly scattered along the amplitude axis showing that the error dependence on A-component is weak. Concerning HPEs, the scatter plot shows A-component values under 1, which means that the scaled average precipitation in the forecast never exceeds three times the observation. ~~On the opposite~~In contrast, A-component negative values are predominant, in particular at LT34, in relation with strong underestimations of the domain-averaged rainfall field. Some cases of significant maximum grid-point errors in conjunction with moderate negative A-component must be related to strong location errors. In these cases, the domain-averaged field may be similar to the observed one while the maximum rainfall is spatially deviated. For the non-HPE~~days~~, we can see that, especially for LT34, the model could significantly overestimate both the A-component and the maximum grid-point error.

The relationship between the different SAL components might help to understand sources of model error. In Fig. 6 the S and A components are drawn for the ~~HPE days~~ HPEs only. Perfect scores are reached for the points located on the origin *O* of the diagram. ~~A very few number of~~ Very few points are located on the top left-hand quadrant. This indicates that an overestimation of precipitation amplitude associated with too small rainfall objects is rarely observed. The points, especially for LT34, are globally oriented from the bottom left-hand corner to the top right-hand corner. This suggests a linear growth of the A-component as a function of the S-component, which means that the average rainfall amount is roughly related to the structure of the spatial

**Table 4.** Pearson correlation between the daily mean S-component and the maximum daily rainfall for the three cluster classifications. A t-test is applied to the individual correlations. For the three clusters, the null hypothesis (true correlation coefficient is equal to zero) is rejected.

| **Cluster** | LT12 | LT34 |
|---|---|---|
| 2 | **0.50** | **0.44** |
| 3 | **0.59** | **0.50** |
| 5 | **0.37** | **0.46** |

extension. For the two diagrams, it can also be noticed that many of the points are situated in the lower-right quadrant, suggesting the presence of too large and/or flat rainfall objects compared to the reference while the corresponding A-component is negative. This is supported by the values of the medians of the ~~two components distribution~~ distribution of the two components(dashed lines) and the quartile values (respective limits of the boxes). ~~This~~ The positive bias in the S-component is even stronger for the most extreme HPEs (red triangles). ~~This~~ The distortion of S-component error compared to A-component shows that the model has more difficulties reproducing the complex spatial structure than simulating the average volume of a heavy rainfall. ~~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~~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).

For each point of the diagram ~~of~~ in Fig. 6 we compute its distance from the origin (perfect score (A=0; S=0)). The dotted circles respectively contain the 25%, 50% and 75% points with the smallest distance. The radius of the circles are much larger for LT34, confirming a degradation of the scores for ~~higher lead time ranges~~longer lead times.

### 3.1.2 Clusters

We use our clustering procedure (as defined in section 2.2.2) to analyze the characteristics of the forecast QPF errors along with the regional properties. SAL components are stated for each day of each cluster associated with HPEs, i.e. C2, C3 and C5. In Fig. 7, PDFs (Probability Density Functions) are drawn from the corresponding normalized histograms for the two lead times LT12 and LT34. The distributions of the A-component are ~~negatively skewed~~ negatively-skewed for all the clusters. This ~~reveals~~ shows that the model tends to produce too weak domain-averaged rainfall in the case of heavy rainfall. This is even more important for clusters 3 and 5. For long lead times, the distributions are flatter, showing that the left tail of the A-component ~~PDFs~~ PDF spreads far away from the perfect score.

The distributions of the S-component (right panels) are positively skewed in cluster 2 and 3, while they are more ~~centred~~ centered for cluster 5. For all the clusters, the spread of the S-component distributions is less dependent on the lead time, compared to the A-component distributions. It is interesting to examine whether a relationship between the S-component and the

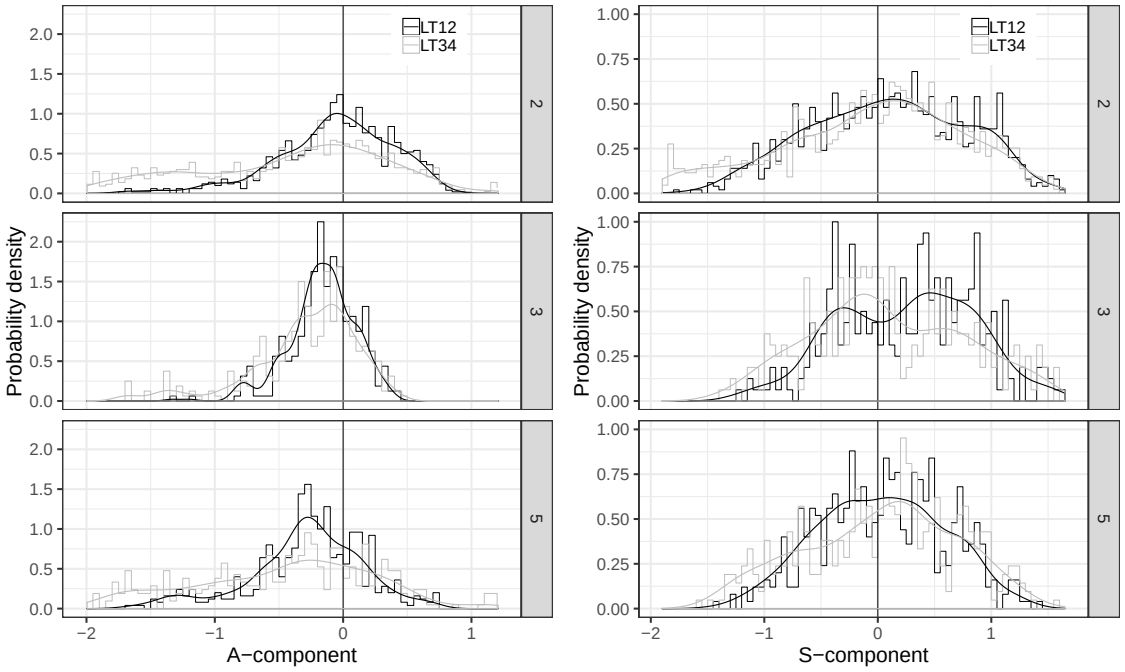


**Figure 7.** A-component (left column) and S-component (right column) normalized histograms and probability density functions for clusters 2, 3 and 5. Results for lead time LT12 are plotted in black lines and results for lead times LT34 are in grey.

intensity of the rainfall ~~is identifiable~~can be identified. A Pearson correlation coefficient is computed between the daily mean of S-component estimated within the ten members of the reforecast and the maximum observed daily rainfall for each cluster class (~~table~~ Table 4). A positive correlation is found for all ~~the three clusters~~three clusters, which corroborates the results from Fig. 6 where HPEs correspond to the highest S-component values. Maximum correlation is found for cluster 3. Although correlations are statistically significant, it is worth noting that values are quite weak (in particular for cluster 5).

### 3.2 Sensitivity to physical parametrizations

The SAL measure is ~~now~~ analysed separately for the ten different physical packages to study corresponding systematic errors. More specifically, we raise the following questions: Do the errors based on an object-quality measure and computed for the different physics implemented in an ensemble system show different rainfall structure properties? Which physical packages are more sensitive to the intense rainfall forecast errors? As in section 3.1, we first distinguish the results for the ~~HPEs~~ HPE group before the cluster ones.

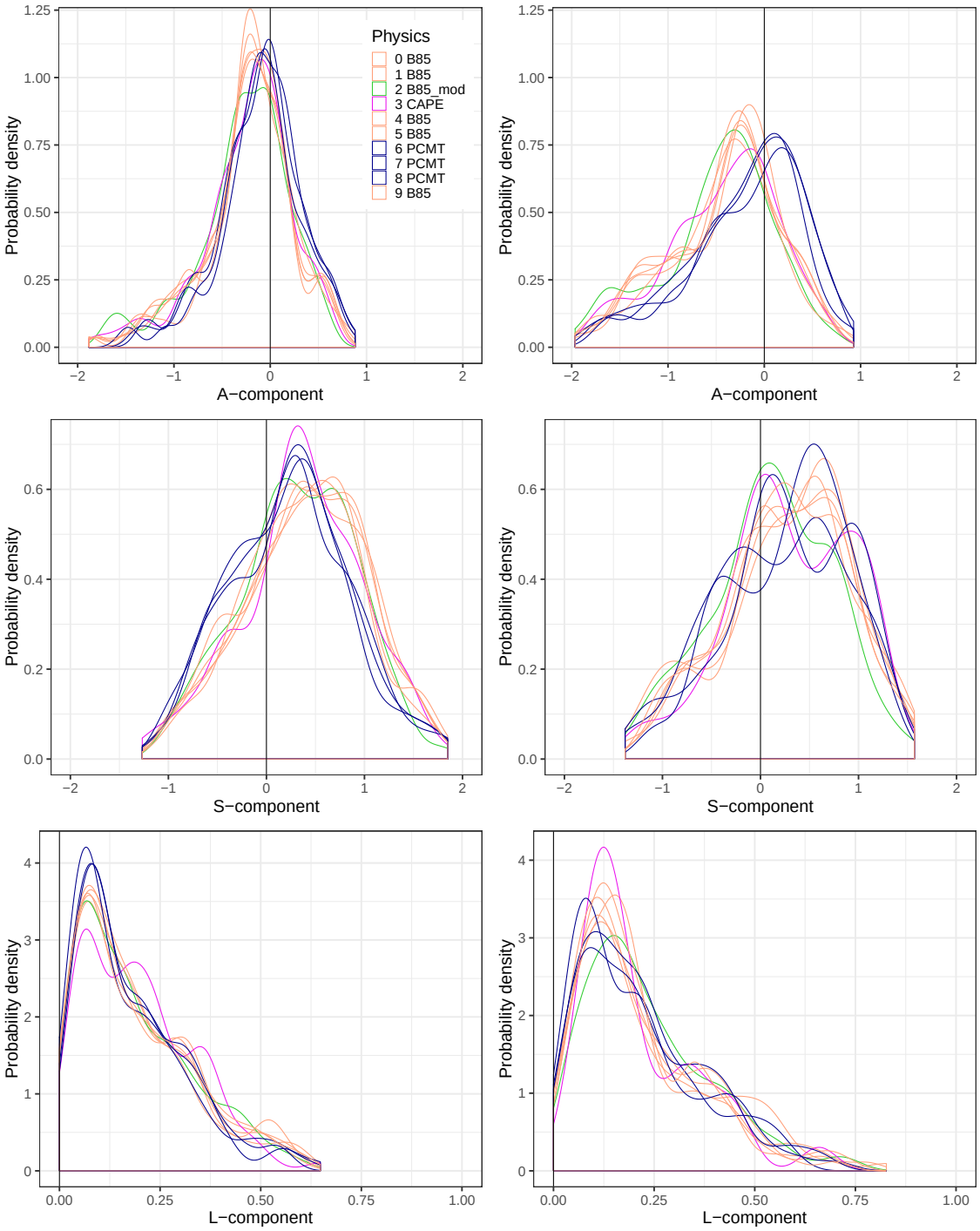


**Figure 8.** Probability density functions of the three SAL components for the HPEs and for each physics of the reforecast system (colored lines). Physics scheme are gathered in four categories depending on the parametrization of the deep convection (PCMT (blue), B85 (orange), $B85_{mod}$ (green), CAPE (purple). Left column corresponds to lead time LT12, and right column relates to lead time LT34.

### 3.2.1 HPEs

Probability density distributions for each SAL component are separately computed for each physics reforecast (Fig. 8), considering only the HPEs~~days~~. Colours lines correspond to four categories, depending on the parametrization of the deep convection. The figure highlights that members from each of the two main parametrization schemes ~~,~~ (B85 and PCMT) have similar behaviours. Considering the A-component, PCMT members are more ~~centred around zero for LT12~~ centered around zero than B85 at LT12. This effect is higher ~~for~~ at LT34, for which B85 and PCMT density distributions are more shifted. ~~For~~ At LT34, more events with a positive A-component are associated ~~to~~ with PCMT, whereas negative values are more recurrent in B85. The A-component never exceeds +1, but ~~strong~~ significant underestimations are observed. This range of values stems from the fact that ~~this~~ the forecast verification is applied to a subsample of the observation limited to the most extreme events. For these specific events, a model underestimation is more frequent than an overestimation. At short lead times, the separation between the two deep convection schemes is also well established for the S-component (middle left panel), but it becomes mixed up ~~for~~ at LT34 (middle right panel). ~~A~~ One reason for this behaviour could be that predictability decreases ~~for~~ at LT34, so that ~~discrepencies of~~ discrepancies in spatial rainfall structure assigned to the physics families become less identifiable. The S-component is positively skewed in all cases (in particular for the B85 physics at LT12 lead time). This supports the previous analysis of the S-component (Fig. 6 and 7), showing that for intense rainfall, the model mostly produces larger and ~~more flat~~ flatter rainfall signal. The results for the S-component also highlight better skills for PCMT schemes ~~for HPEs events~~than for HPEs, especially at ~~first~~ short lead times. Focusing on high values of S, B85 exhibits a stronger distribution tail at LT12, while both schemes seem comparable for LT34.

For the L-component, the maxima of the density distributions are higher for PCMT ~~than B85 for~~ at lead time LT12, implying a more significant number of good estimations of pattern location. Regarding the tail of the L-component PDF, it is globally more pronounced ~~in~~ at LT34 than LT12. This means that the location of HPEs is poorly forecasted at long lead times. Concerning the behaviour of the forecasts that use the CAPE or $B85_{mod}$ schemes, their A-component PDFs are close to the B85 PDFs. This is not observed for the other components. For the S-component, the CAPE distribution follows ~~, at LT12,~~ the PCMT one at LT12. For the L-component, $B85_{mod}$ PDF is close to the B85 ones, while CAPE shows ~~a behaviour different~~ different behaviour from all the ~~others physics.~~ other physics. The use of a closure based on CAPE, rather than on the convergence of humidity seems to modulate the location of precipitation produced by this deep convection parametrization scheme. Moreover, at LT34 CAPE is characterized by a lower number of strong location errors, compared to the other physics.

### 3.2.2 Clusters

According to the results of the previous section, which ~~shows~~ show that the predictability of intense rainfall events is sensitive to the parametrization of the deep convection, we ~~continue analysing the~~ have continued to analyze the model behaviour for the four different deep convection schemes~~model behaviours~~: B85, $B85_{mod}$, CAPE, and PCMT. The link between the behaviour of the physical schemes and ~~the~~ belonging to a particular cluster is statistically assessed through the SAL ~~components~~ component differences between the schemes.

**Table 5.** This table is removed~~The table provides the $p$-values computed from the $k$-sample AD test at 0.05 significance level for LT12. Upper right side of the table displays the values for the A-component obtained for pairs of distributions from the four physical package families. The elements between brackets represent the $p$-values computed for each of the three clusters retained after the clusterization. The lower left side shows results for the S-component. Bold values indicates where the null hypothesis is rejected, meaning that the difference between two distributions is statistically significant.~~

| **Physics** | B85 | $B85_{mod}$ | CAPE | PCMT |
|---|---|---|---|---|
| B85 | / | (0.77)(0.46)(0.65) | (0.83)(0.74)(0.56) | (0.16)(**0.00**)(**0.02**) |
| $B85_{mod}$ | (0.24)(0.39)(0.60) | / | (0.99)(0.99)(0.78) | (0.14)(**0.01**)(0.13) |
| CAPE | (0.25)(0.80)(0.74) | (0.33)(0.66)(0.45) | / | (0.23)(**0.02**)(0.70) |
| PCMT | (**0.05**)(**0.02**)(**0.01**) | (0.61)(0.72)(0.38) | (0.13)(0.39)(**0.02**) | / |

Any parametric goodness-of-fit tests, which ~~assumes~~ assume normality, have been discarded, because SAL values are not normally distributed. We choose the $k$-sample Anderson–Darling (AD) test (Scholz and Stephens, 1987; Mittermaier et al., 2015), in order to evaluate whether differences between two given distributions are statistically significant. It is an extension of the two-sample test (Darling, 1957), originally developed starting from the Classic Anderson-Darling test (Anderson and Darling, 1952). The $k$-sample AD test is a non parametric test designed to compare continuous or discrete sub-samples of the same distribution. In this case the test is implemented for the evaluation of the pairs of distributions. ~~The two sample goodness-of-fit statistic $A^2_{mn}$ is a sum of the integrated squared differences between two distributions functions:~~

$$A^2_{mn} = \frac{mn}{N} \int_{-\infty}^{+\infty} \frac{\{F_m(x) - G_n(x)\}^2}{H_N(x)\{1 - H_N(x)\}} dH_N(x)$$

~~where $F_m(x)$ is the proportion of the sample $X_1, \ldots, X_m$ that is not greater than $x$ and $G_n(x)$ is the empirical distribution function of the second independent sample $Y_1, \ldots, Y_n$ obtained from a continuous population with distribution function $G(x)$ and $H_N(x) = \{(mF_M(x) + nG_n(x))/N$ , with $N = m + n\}$ is the empirical distribution function of the pooled sample. Since $n$ can differ from $m$, the test does not require samples with the same size. The above integrand is appropriately defined to be zero whenever $H_N(x) + 1$ is equal to zero. Under the null hypothesis $H_0$, for which $F(x) = G(x)$, the expected value of $A^2_{mn}$ is 1. The test statistic is *standardized* using the expected value and the variance of $A^2_{mn}$, $\sigma_N$:~~

$$T_N = \frac{A^2_{mn} - 1}{\sigma_N}$$

~~The null hypothesis $H_0$ is rejected if $T_N$ exceeds the critical value $t(\alpha)$, where $\alpha$ is the significance level, here set to 0.05. If this condition is verified, distributions are significantly different from each other at the 5% level.~~

The tests are performed for the comparison of each ~~pairs~~ pair of PDFs combined from the four deep convection families and from the three clusters ~~classifications (Tab. 5 (LT12) and Tab. 6 (LT34)). Statistically significant differences are found for A-components for both LT12 and LT34. For the most intense events, $B85_{mod}$ and CAPE perform as B85 (see $p$-values for~~

**Table 6.** This table is removed~~As in Tab. 5, but for lead time LT34.~~

| **Physics** | B85 | $B85_{mod}$ | CAPE | PCMT |
|---|---|---|---|---|
| B85 | / | (0.16)(0.25)(0.91) | (0.45)(0.28)(0.13) | (**0.01**)(**0.00**)(**0.00**) |
| $B85_{mod}$ | (0.23)(0.73)(0.40) | / | (0.92)(1.00)(0.40) | (**0.00**)(**0.01**)(0.08) |
| CAPE | (0.84)(0.77)(0.35) | (0.64)(0.93)(0.28) | / | (**0.01**)(**0.02**)(0.79) |
| PCMT | (0.24)(0.37)(0.50) | (0.63)(0.81)(0.77) | (0.50)(0.49)(0.47) | / |

~~cluster 3 in the upper side of Tab. 5 and 6). It is worth noting that when comparing $B85_{mod}$ to CAPE, $p$-values are close to 1, so that these parametrizations exhibits the same mean behaviour for the~~ classification. For the A-component ~~over the full dataset. As already observed in the HPEs analysis section,~~ PCMT physics distributions depart significantly from B85 schemes at all lead times~~. PCMT,~~, while $B85_{mod}$ ~~and CAPE exhibit differences for some clusters, especially at LT34. By focusing the attention on the most extreme clusters, it is worth noting that PCMT,~~ and CAPE perform as $B85_{mod}$ ~~and CAPE distributions are equivalent for cluster 5, but not for cluster 3. This means that~~, meaning that the modified versions of B85 $_{mod}$ ~~and CAPE are alternatively close to B85 or PCMT, depending on the cluster.~~ weakly affect physics behaviour (not shown).

~~In~~ With respect to the S-component distributions, $k$-sample AD tests show significant differences between B85 and PCMT physics for LT12~~. For longer~~, but not for the longest lead times (~~LT34)structure quality measures converge towards a homogeneous distribution~~ not shown). At LT34 we observe a convergence of the physics ~~schemes (see $p$-values in the bottom side of Tab. 6)~~scheme towards a homogeneous distribution towards a homogeneous distribution, meaning that the differences between physics are negligible.

The test applied to the location component ~~(not shown)~~ does not reveal significant differences between the PDFs. We suppose that the limited dimensions of the domain employed in this study, as well as its irregular shape, may lead to a less coherent estimation of the location, resulting in a degradation of the score significance. Since the L-component result is not informative about HPEs, it is ignored hereafter.

Once the statistical differences between the PDFs of the physics have been examined, it is interesting to compare the relative error on the amplitude and structure components. S and A component errors are estimated by comparing the shapes of their distributions. Empirical Cumulative Density Functions (ECDF) of S and A components are computed separately for each cluster and lead time (LT12 and LT34). We show an example of an ECDF for cluster 2 ~~and~~ at LT34 (Fig. 9). Forecasts are perfect when the ECDF tends towards ~~an~~ a Heaviside step function, which means that the distribution tends towards the Dirac delta function ~~centred~~ centered on zero. ~~The departure~~ These functions are estimated over a bounded interval, corresponding to the finite range of S and A components. The deviation from the perfect score ~~could be~~ was quantified, by estimating the area under the ECDF curve on the left side, and the area above the ECDF curve on the right side:

$$err_{-} = \int_{-2}^{0} F(tx)dt - dx - \int_{-2}^{0} H(x)dx = \int_{-2}^{0} F(tx)dt - 0dx - 0 = \int_{-2}^{0} F(tx)dtdx, \quad (9)$$

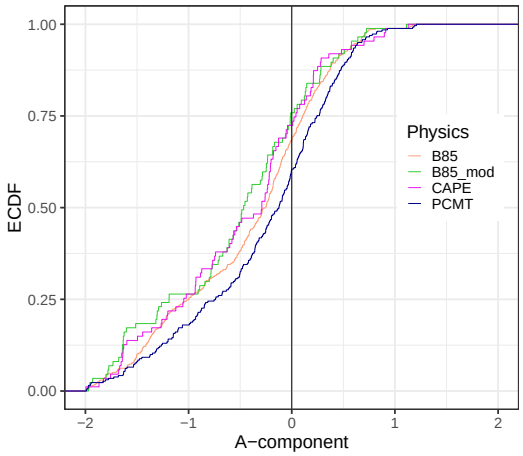


**Figure 9.** Empirical cumulative distribution function of ~~of~~ the A-component computed from ~~the~~ cluster 2 ~~and~~ at lead time LT34 for the four classes of physics schemes.

**Table 7.** This table is removed~~Forecast errors computed from the A-component distribution for the B85 and PCMT physics for each cluster classifications and lead time LT12. Grey lines denote clusters for which B85 and PCMT differences are not statistically significant.~~

| | B85 | | | PCMT | | |
|---|---|---|---|---|---|---|
| Cluster | $A_-$ | $A_+$ | $A_{tot}$ | $A_-$ | $A_+$ | $A_{tot}$ |
| 2 | 0.216 | 0.136 | 0.352 | 0.185 | 0.163 | 0.347 |
| 3 | 0.215 | 0.029 | 0.244 | 0.140 | 0.054 | 0.194 |
| 5 | 0.369 | 0.045 | 0.413 | 0.295 | 0.075 | 0.370 |

$$err_+ = \int_0^2 H(x)dx - \int_0^2 F(x)dx = 2 - \int_0^2 F(x)dx, \tag{10}$$

$$err = err_- + err_+ = 2 - \int_0^2 F(x)dx + \int_{-2}^0 F(tx)dtdx, \tag{11}$$

where $F(x)$ is the ECDF computed for A or S, $H(x)$ is the Heaviside step function and $err$ is the forecast error for a given component. The lower and upper boundaries of the integrals are equal to -2 and +2, because A and S components range between these two values by construction. Since the previous $k$-sample AD test ~~pointed out~~ highlighted significant differences within the two main classes B85 and PCMT, the evaluation of the errors is ~~restrained~~ limited to these two specific classes.

~~The results for~~ The results of the error diagnostic $err$ for the the A-component are shown in ~~Table 7 (LT12) and Table 8 (LT34). Negative and positive errors~~ Fig. 10(a). Errors increase with lead time. We note that the negative errors are always

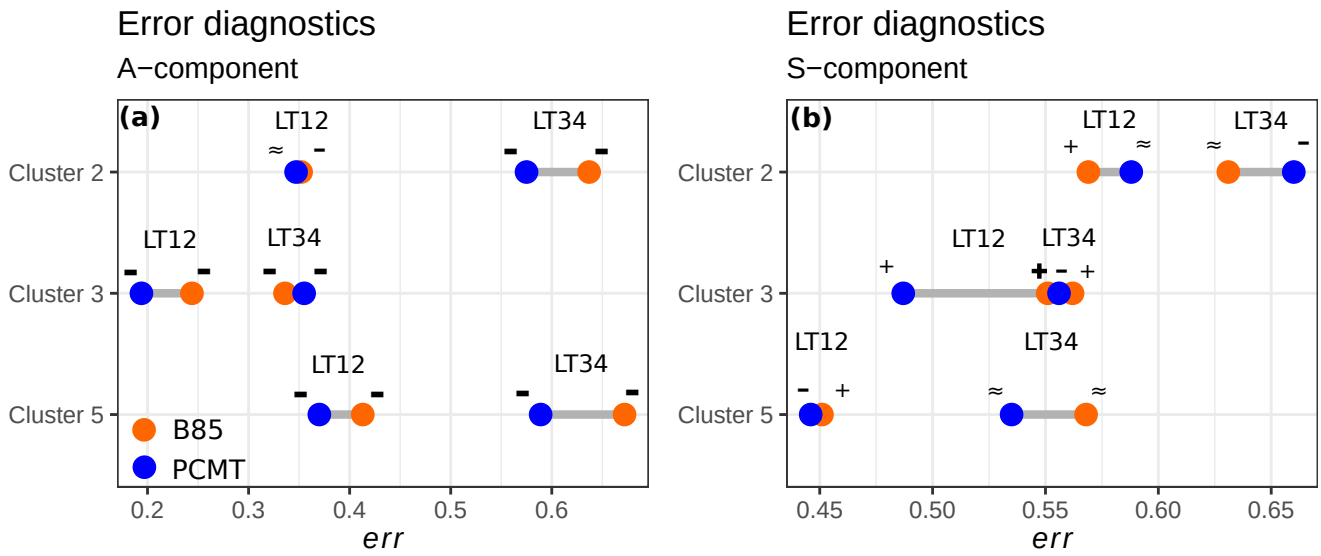


**Figure 10.** ~~–~~This figure is added.~~–~~ Dumbbell plot of integrated error diagnostics computed using eq. 11. Colours refer to B85 (orange) and PCMT (blue) deep convection parametrization schemes. Results are stratified on the basis of the clusters and lead times. Symbols denote whether positive or negative errors dominate. These signs are defined using the following definition: − (bold) if $\frac{err_-}{err_+} \geq 2$; − if $1.1 \leq \frac{err_-}{err_+} < 2$; ≈ if $0.9 < \frac{err_-}{err_+} < 1.1$; + if $0.5 < \frac{err_-}{err_+} \leq 0.9$; + (bold) if $\frac{err_-}{err_+} \leq 0.5$.

**Table 8.** This table is removed~~As in Table 7, but for lead time LT34.~~

| | B85 | | | PCMT | | |
|---|---|---|---|---|---|---|
| Cluster | $A_-$ | $A_+$ | $A_{tot}$ | $A_-$ | $A_+$ | $A_{tot}$ |
| 2 | 0.530 | 0.107 | 0.637 | 0.429 | 0.146 | 0.575 |
| 3 | 0.312 | 0.023 | 0.336 | 0.294 | 0.061 | 0.355 |
| 5 | 0.608 | 0.064 | 0.672 | 0.451 | 0.138 | 0.589 |

**Table 9.** This table is removed~~As in Table 7, but for the S-component.~~

| | B85 | | | PCMT | | |
|---|---|---|---|---|---|---|
| Cluster | $S_-$ | $S_+$ | $S_{tot}$ | $S_-$ | $S_+$ | $S_{tot}$ |
| 2 | 0.230 | 0.338 | 0.569 | 0.285 | 0.303 | 0.588 |
| 3 | 0.128 | 0.434 | 0.562 | 0.171 | 0.316 | 0.487 |
| 5 | 0.187 | 0.264 | 0.451 | 0.268 | 0.178 | 0.446 |

**Table 10.** This table is removedAs in Table 8, but for the S-component.

| Cluster | B85 | | | PCMT | | |
|---|---|---|---|---|---|---|
| | $S_-$ | $S_+$ | $S_{tot}$ | $S_-$ | $S_+$ | $S_{tot}$ |
| 2 | 0.326 | 0.305 | 0.631 | 0.381 | 0.279 | 0.660 |
| 3 | 0.214 | 0.337 | 0.551 | 0.246 | 0.310 | 0.556 |
| 5 | 0.278 | 0.289 | 0.568 | 0.269 | 0.265 | 0.535 |

more important than at least twice as large as the positive ones. This behaviour is strengthened at LT34, especially for clusters 3 and 5. This is not surprising since those two clusters collect the most extreme rainfall events. Indeed, the uncertainty of the forecast is supposed to be higher in the case of most intense rainfall events. Forecast hardly produce as many rainfall amount as it is observed, especially for the longest lead times, since temporal error can lead to strong underestimations.

PCMT produces overall better Forecasted averaged rainfall amounts are almost always underestimated. PCMT produces overall better A-component statistics, in particular the A-component negative contribution is reducedstatistics, except for cluster 3 at LT34. It is interesting to observe that the weakest errors are associated with cluster 3, which is the most extreme one. Since cluster 3 collects a large number of precipitation events impacting the Cévennes chain, we may suppose that the domain averaged rainfall amounts are more predictable in situations of precipitation driven by the orography. Concerning the S-component evaluation , results are shown in Tables 9 and 10. The best forecasts are observed for clusters 3 and 5. This suggests that the forecast of the structure of the object depends on the considered phenomenon. In particular, neglecting the location error, shape and size (see Fig. 10(b)), structures of rainfall patterns are better forecasted for heavy rainfall events (clusters 3 and 5), rather than for the remaining classes of events. In contrast to the A-component, the S-component exhibits the highest error on the right side of the distribution $err_+$ for B85 scheme for most of the cases (majority of + sign in Fig. 10(b)), whereas this trend is not systematic in for PCMT physics. Restraining the analysis to LT12 PCMT globally performs better than B85, as the positive bias of the S-component is reduced. except for cluster 2. As with the amplitude A, the S-component gets worse for longer lead timetimes, resulting in a shift to more negative values of the distribution larger $err_-$ for both B85 and PCMT physics . (more - sign for LT34 in Fig. 10(a) and (b)). The lowest values of S-component are achieved for cluster 5. Cluster 5 HPEs are known to have specific regional properties whose influence on S-component results should be studied with further diagnostics.

### 3.2.3 Rainfall object analysis

We now analyze the physical properties of the objects, i.e. the number of objects from a rainfall field , and the object integrated volumes and surfaces, according to the different clusters. All the statistics are applied separately to the B85, PCMT physics, and observations. For each day of the dataset period, the thresholds defined in subsection 2.3.1 lead to the identification of a certain number of precipitating objects. The frequency of this the number of objects per day is plotted by means of normalized histograms for the three clusters (Fig. 11). Cluster Clusters 2 and 3 show maximum frequency for one and three objects

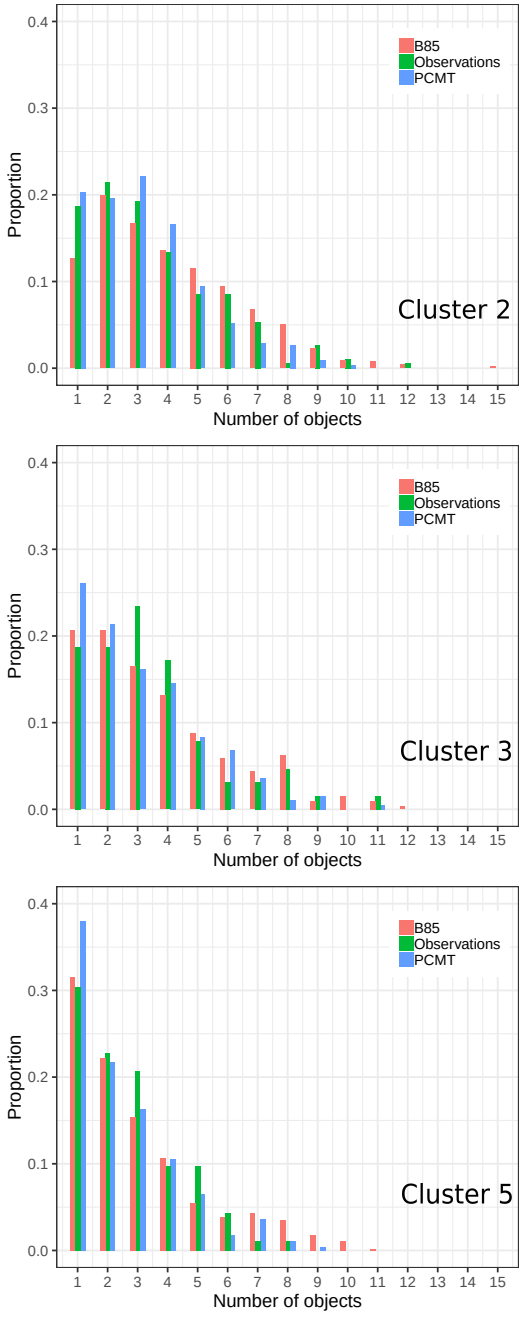


**Figure 11.** Normalized histograms of the daily number of SAL patterns, for B85 physics scheme (red), PCMT (blue), Observation (green). Panels correspond to the 3 clusters classification.

object range, whereas cluster 5 is dominated by one object per day. ~~A visual inspection of the individual cases of single object rainfall patterns in~~ This specific property of cluster 5 ~~suggests that the zone of the domain affected by objects can be crucial. Single objects extend mainly over the Languedoc-Roussillon region (cluster 5), while the other clusters~~ can explain the best result obtained for S-component (section 3.2.2). Indeed, we may assume that S-component estimation is more accurate for a one-to-one object comparison. The other clusters frequently display rainfall accumulated bands ~~frequently~~ split over the domain, typically over the Cévennes and Alpine regions. ~~Objects~~ Object identification for PCMT forecast shows ~~the existence of~~ that there is an overestimation of single ~~objects~~ object days compared to the observation and to B85 physics scheme, a behaviour emphasized in clusters 3 and 5.

More details about the magnitude of the objects can be ~~achieved~~ produced by computing the integrated mass per object, $M_k$ (see subsection 2.3.1). First, for each day, objects are sorted from the largest to the smallest integrated mass. Integrated mass distribution of the two *heaviest* objects (noted $O_1$ and $O_2$) are then dispatched as a function of the number of objects for each cluster on Fig. 12. First, the range value of M is highly variable from one cluster to another. Maximum values are observed for cluster 3, while the magnitude for clusters 2 and 5 is comparable. The decrease of the mass for $O_1$ ~~M is more clear~~ is clearer for cluster 3, meaning that a high number of objects over the domain leads to a natural decrease of the M value of the heaviest ones. We think that a part of the total integrated mass is then redistributed to the other objects. This is confirmed ~~on~~ by $O_2$ curves since its mass increases with the number of objects. Conversely, for cluster 5, $O_1$ mass increases with the number of the objects, while $O_2$ is almost stable. The gap between $O_1$ and $O_2$ masses is maximum in the most extreme clusters (3 and 5). This suggests that when computing the volume V (see eq. ~~()~~ 7) and L2 (see ~~()~~eq. 4), the weighted average is dominated by the object $O_1$. This implies that the verification could be considered as a single to single object metric.

~~The integrated mass M is only partially informative about the intensity of accumulated rainfall because it depends also on the spatial extension of object, also called the object base area. We define as $R^*$ the integrated individual object mass $M$, weighted by its base area. The same statistics than previously are shown for $R^*$ in Fig. 13. Compared to $M$, the gap between $O_1$ and $O_2$ is significantly reduced for $R^*$, even if $R^*$ is still larger for $O_1$ than for $O_2$. $R^*$ reaches the greatest values for cluster 3, while, in contrast with the results from $M$, cluster 5 exhibits higher $R^*$ compared to cluster 2. This difference is explained by considering the object base area values. Pattern spatial extension are frequently larger for cluster 2, than for cluster 5. Orography leads cluster 2 to have more spatially extended objects with a weaker scaled object mass $R^*$ than those of cluster 3.~~

~~The clusters associated with rainfall events impacting the Cevennes and eastern area of the domain D (clusters 3 and 5) are characterized by similar values of base area (not shown). Accordingly, they collect similar phenomena, but for two distinct classes of intensities. It can also be noted for cluster 5 that $R^*$ is slightly decreasing, meaning that base area values increase faster than integrated mass values per number of identified objects.~~

~~Comparing observed and forecast objects we can see that the scaled pattern mass criterion highlights the gap between observations and models for $O_1$ and $O_2$, especially for clusters 3 and 5. B85 physics usually underestimates $R^*$ compared to the observations except for the highest number of objects. On the contrary, for PCMT the departures between models and~~

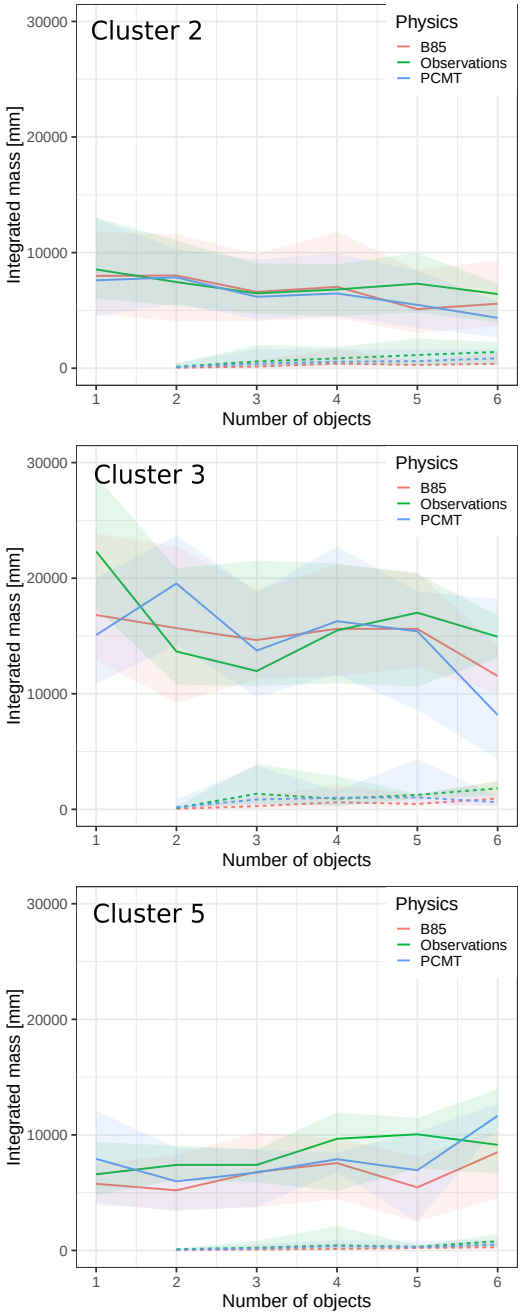


**Figure 12.** Distribution of SAL first pattern $O_1$ rain amount according to the number of patterns per day. Curves stand for the median of the distribution, shaded areas range between $25\%$ and $75\%$ percentiles. The dashed lines correspond to the second ranked SAL pattern $O_2$ rain amount.

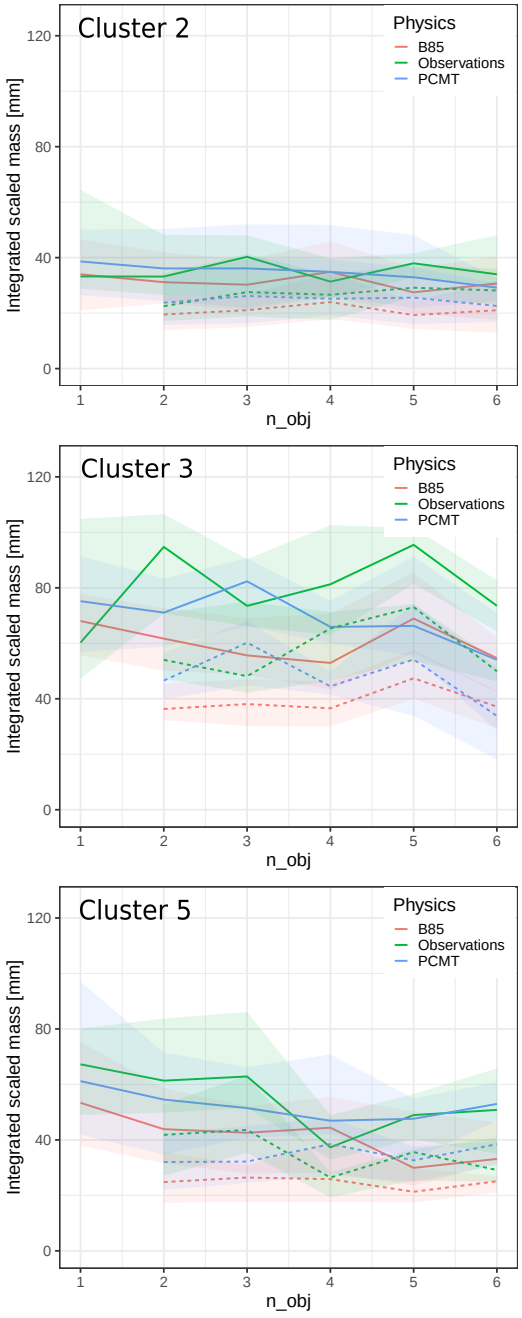


**Figure 13.** This figure is removed~~As in Fig. 12, but for rain amounts scaled by the pattern base area surface.~~

~~observations for $R^*$ are higher in the most extreme clusters (3 and 5), showing a relation between the error and the magnitude of the observed variable.~~

We now examine the ratio between the daily maximum rainfall of objects $O_1$ and $O_2$. This ratio ranges between 1.5 and 3 which means that $O_1$ represents the essential contribution of the daily rainfall peak~~, even when its scaled object mass $R^*$ is close to $O_7$~~. Since $O_1$ base area tends to be significantly larger that $O_2$, the information related to the inner object maximum rainfall is diluted in the large base area, resulting in a flat weak mean intensity of the object. This last result appears to support the fact that SAL metric gives more weight to the object that contains the most intense rainfall.

The comparison between the model reforecast physics and the observations is addressed using the whole distribution of daily mass $M$ from the objects $O_i$ identified across the full reforecast dataset, where $i$ ranges between 1 and the total number N of objects. We proceed separately for each physical package. For a given scheme and cluster, the quantile values corresponding to the selected dataset are sorted in ascending order, and then plotted versus the quantiles calculated from observations (Fig. 14). Half of the quantile distributions are not visible as they correspond to very weak pattern masses. For cluster 2 and PCMT physics most of the distribution of object mass is close to the observations, however all ~~the~~ other physics distributions are skewed to the right compared to the observations for values below 10000. This behaviour is also observed for cluster 5 and it involves PCMT physics as well, for values between percentile $0.5$ and percentile $0.7$. Overall, in the quantile-quantile plot for cluster 5, the PCMT outperforms B85. In cluster 3, discrepancies between PCMT, B85 and the observations are of opposite sign, with PCMT being slightly above the observations, while B85 showing a weak underestimation. CAPE physics distribution is left skewed compared to the observations and to the ~~others~~ other physics. These results ~~reveal~~ highlight some interesting properties of the models in predicting the rainfall objects. ~~For the most extreme clusters~~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 ~~5.~~ 3. This means that the forecast is able to reproduce the same proportion of rainfall amounts inside a feature as the observations, even concerning the extreme right tail of the distributions, which corresponds to the ~~the~~ major events of the series.

## 4 Summary and conclusions

In this study we have characterized the systematic errors of 24-hour rainfall amounts from a reforecast ensemble dataset, covering a 30-year fall period. A 24-hour rainfall observation reference has been produced ~~with the same model resolution as the reforecast one~~ on a regular grid with a resolution identical to the model in order to ~~have access to a~~ run point-to-point verification. We applied an object-based quality measure in order to evaluate the performance of the forecasts of any kind of ~~HPEs~~HPE. Then, we ~~take~~ took advantage of a rainfall clustering to ~~analyse~~ analyze the dependence of systematic errors ~~to~~ on clusters.

The selection of the HPEs within the reference dataset was based on a peak-over-threshold approach. The spatial regional discrepancies between HPEs are ~~studied based on the~~ highlighted on the basis of the $k$-means clustering of the 24-hour rainfall.

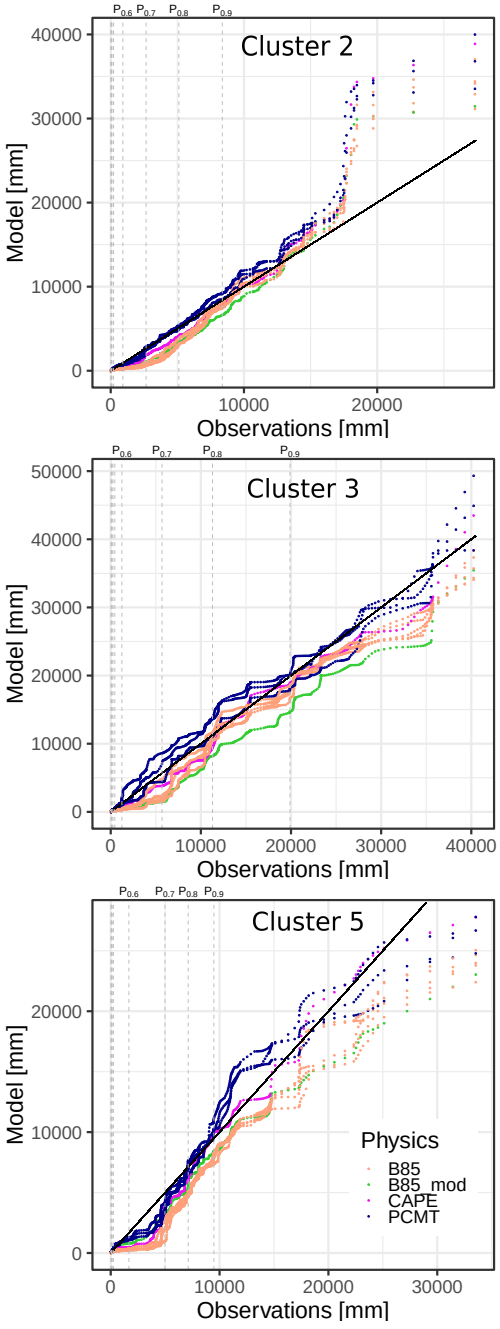


**Figure 14.** Quantile-quantile plot between SAL pattern rain amounts from the model (Y-axis) and from the observation (X-axis). Physics schemes are gathered into 4 classes (B85, PCMT, B85mod, CAPE). Observation deciles correspond to the vertical dashed lines.

Finally, we ~~analysed the rainfall objects properties repectively~~ analyzed the rainfall object properties respectively in the model and in the observation to underline the rainfall field object properties for which the model acts distinctly.

The peak-over-threshold criterion ~~leads~~ lead to the selection of 192 HPEs, confirming that the most impacted regions are the Cévennes area and part of the Alps. ~~Even though HPEs affects predominantly the mountainous areas, severe precipitating systems can occur in plain areas, especially on the foothills oriented towards the meridional fluxes. The~~ The composite analysis for the five clusters ~~reveals~~ shows that each cluster is associated ~~to~~ with a specific class and location of 24-hour precipitation events. It was found that 86% of the ~~total~~ number of HPEs are included in clusters 2, 3 and 5. Cluster 2 and 3 ~~patterns impact predominantly~~ HPEs predominantly impact the Cévennes and Alps area, while ~~the~~ cluster 5 HPEs are located over the Languedoc-Roussillon region. Moreover clusters 3 and 5 ~~are~~ include the most extreme ones~~, while cluster 4 contains weak rainfall events or dry days. Diagnostics~~ . Only diagnostics for clusters 2, 3 and 5 ~~only~~ are considered.

~~Model performances analysis have lead to several distinct results that we outline in the following.~~

The SAL object-quality measure has been applied distinctly to the ten ~~model physics members~~ physics schemes (one per member) of the reforecast dataset and compared to the rainfall reference. ~~This~~ It shows that the model~~overall behaviour~~ 's overall behaviour for HPE forecasting is characterized by negative A-components and positive S-components. ~~The model objects are generally more flat and large objects than the observed ones, and moreover~~ 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~~. For all computed performance diagnostics, it has been found a degradation of SAL scores along with the lead times, comparatively with quantitative rainfall diagnostics~~, it does not 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.

~~When SAL diagnostics are performed~~ In order to show regional disparities in the model behaviour, the SAL diagnostics have been divided according to the clusters ~~, the A-component is negative-skewed, and it is enhanced notably~~ 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 Amplitude SAL-component. It is notably the case for the most extreme clusters (over the Cévennes and over the Languedoc-Roussillon). ~~Concerning the structure~~ 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 ~~behaviour, diagnostic is dependent to the clusters. It is~~ 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 ~~centred. This might indicate that heavy rainfall episodes over the relief regions (Cévennes, Alps) are represented by the model by flat and large pattern spreading out on a larger zone compared to the observations. For cluster~~ 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, ~~this effect is not found, and at that point~~ for which the S-component reaches the best score. Though it is difficult at this point to determine whether this ~~is characterising more a~~

characterizes an actual contrast in the model behaviour or ~~whether~~ 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 ~~performances of the model are then investigatedseparately for each physical scheme composing the reforecast dataset, emphasising mostly~~ impact of the different physics schemes has also been investigated, and it mostly emphasized the role of the deep convection physical ~~parametrisation. In terms of SAL~~ parameterization. Considering the SAL diagnostics, the two main deep convection schemes, B85 and PCMT, clearly determine the behaviour of the model ~~. However, for~~ in HPE forecasting until lead time ranges ~~higher~~ longer than three days, after which no significant differences appear. ~~It has been measured quantitatively that PCMT members~~ This difference is clearly in favour of the PCMT scheme which performs better than B85 ~~ones in terms of both SAL diagnostics,~~ for both SAL A and S components ~~. S-component analysis shows to be better also for HPEs rather than for weak or moderate events which means that~~ 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 ~~predictability of pattern structure is higher for HPEs~~S-component for the most intense rainfall, which shows that PCMT better approximates the structure of the rainfall patterns in these cases.

~~The second part of the study was dedicated to the characterization of rainfall objects properties in the model and in the reference, cluster by cluster. Cluster~~ 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, ~~which depicts essentially the precipitating objects that impacts the~~ characterized by the Languedoc-Roussillon ~~, is the only cluster characterized by~~ HPEs, the rainfall distribution could even be considered as a single object rainfall field. ~~The analysis of object masses distribution of the two first sorted objects ($O_1$, $O_2$), shows that the second ranked object weight is weaker also in term of inner rainfall maximum which means that the weight of the larger object $O_1$ is preponderant in the SAL-analysis.~~

~~The analysis of the ranked~~ Then we focused on the ranked distributions (quantile-quantile analysis) of the object masses ~~shows that weakest precipitations~~ 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. ~~On another hand,~~ However, looking at the object mass distributions for the whole period, we find they are relatively close between all the physics ~~scheme~~ schemes and the observation for most extreme rainfall events, ~~specially~~ 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 one 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 focusing on the study of 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.

The inter-comparison between some model physics deep convection ~~scheme~~ schemes and their role in HPEs predictability shows it is of course very sensitive for designing multi-physics type of ensemble forecasting systems. ~~Even if~~ While the sensitivity to the initial perturbations was not studied in this work, the forecast of intense rainfall seems to be mainly driven by the classes of deep convection parametrizations. Since physical parametrization set-up is built by replicated schemes, the model error representation might lack ~~of~~ an exhaustive sampling of the forecasted trajectories. Using more than two deep convection parametrization schemes may improve the representation of model errors, at least for heavy ~~precipitating~~ precipitation events.

*Data availability.* Research data can be accessed by contacting Matteo Ponzano at his e-mail address matteo.ponzano@meteo.fr and the other authors.

*Author contributions.* MP, BJ, and LD conceived and designed the study. MP carried out the formal analysis, wrote the whole paper, made the literature review, and produced the observation reference dataset. BJ built the hindcast dataset. BJ, LD, and PA reviewed and edited the original draft.

*Competing interests.* The authors declare that they have no conflict of interest.

## References


AghaKouchak, A., Behrangi, A., Sorooshian, S., Hsu, K., and Amitai, E.: Evaluation of Satellite-Retrieved Extreme Precipitation Rates across the Central United States, Journal of Geophysical Research: Atmospheres, 116, https://doi.org/10.1029/2010JD014741, 2011.

Anagnostopoulou, C. and Tolika, K.: Extreme Precipitation in Europe: Statistical Threshold Selection Based on Climatological Criteria, Theoretical and Applied Climatology, 107, 479–489, https://doi.org/10.1007/s00704-011-0487-8, 2012.

Anderson, T. W. and Darling, D. A.: Asymptotic Theory of Certain "Goodness of Fit" Criteria Based on Stochastic Processes, The Annals of Mathematical Statistics, 23, 193–212, https://doi.org/10.1214/aoms/1177729437, 1952.

Argence, S., Lambert, D., Richard, E., Chaboureau, J.-P., and Söhne, N.: 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, 1775–1788, https://doi.org/10.1002/qj.314, 2008.

Bazile, E., Marquet, P., Bouteloup, Y., and Bouyssel, F.: The Turbulent Kinetic Energy (TKE) scheme in the NWP models at Meteo France, in: Workshop on Workshop on Diurnal cycles and the stable boundary layer, 7-10 November 2011, pp. 127–135, ECMWF, ECMWF, Shinfield Park, Reading, 2012.

Bechtold, P., Bazile, E., Guichard, F., Mascart, P., and Richard, E.: A Mass-Flux Convection Scheme for Regional and Global Models, Quarterly Journal of the Royal Meteorological Society, 127, 869–886, https://doi.org/10.1002/qj.49712757309, 2001.

Belamari, S.: Report on uncertainty estimates of an optimal bulk formulation for surface turbulent fluxes, MERSEA IP Deliverable 412, pp. 1–29, 2005.

Berner, J., Shutts, G. J., Leutbecher, M., and Palmer, T. N.: A Spectral Stochastic Kinetic Energy Backscatter Scheme and Its Impact on Flow-Dependent Predictability in the ECMWF Ensemble Prediction System, Journal of the Atmospheric Sciences, 66, 603–626, https://doi.org/10.1175/2008JAS2677.1, 2009.

Boisserie, M., Descamps, L., and Arbogast, P.: Calibrated Forecasts of Extreme Windstorms Using the Extreme Forecast Index (EFI) and Shift of Tails (SOT), Weather and Forecasting, 31, 1573–1589, https://doi.org/10.1175/WAF-D-15-0027.1, 2015.

Boisserie, M., Decharme, B., Descamps, L., and Arbogast, P.: Land surface initialization strategy for a global reforecast dataset, Quarterly Journal of the Royal Meteorological Society, 142, 880–888, https://doi.org/10.1002/qj.2688, http://rmets.onlinelibrary.wiley.com/doi/full/10.1002/qj.2688, 2016.

Bougeault, P.: A Simple Parameterization of the Large-Scale Effects of Cumulus Convection, Monthly Weather Review, 113, 2108–2121, https://doi.org/10.1175/1520-0493(1985)113<2108:ASPOTL>2.0.CO;2, 1985.

Buizza, R. and Palmer, T. N.: The Singular-Vector Structure of the Atmospheric Global Circulation, Journal of the Atmospheric Sciences, 52, 1434–1456, https://doi.org/10.1175/1520-0469(1995)052<1434:TSVSOT>2.0.CO;2, 1995.

Caldas-Álvarez, A., Khodayar, S., and Bock, O.: 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, https://doi.org/10.5194/asr-14-157-2017, https://www.adv-sci-res.net/14/157/2017/, 2017.

Charron, M., Pellerin, G., Spacek, L., Houtekamer, P. L., Gagnon, N., Mitchell, H. L., and Michelin, L.: Toward Random Sampling of Model Error in the Canadian Ensemble Prediction System, Monthly Weather Review, 138, 1877–1901, https://doi.org/10.1175/2009MWR3187.1, 2009.

Collier, C. G.: Flash Flood Forecasting: What Are the Limits of Predictability?, Quarterly Journal of the Royal Meteorological Society, 133, 3–23, https://doi.org/10.1002/qj.29, 2007.

Courtier, P., Freydier, C., Geleyn, J., Rabier, F., and Rochas, M.: The ARPEGE project at Météo-France, ECMWF Seminar proceedings, vol. II. ECMWF Reading, UK, pp. 193–231, 1991.

Cuxart, J., Bougeault, P., and Redelsperger, J.-L.: A turbulence scheme allowing for mesoscale and large-eddy simulations, Quarterly Journal of the Royal Meteorological Society, 126, 1–30, https://doi.org/10.1002/qj.49712656202, 2000.

Darling, D. A.: The Kolmogorov-Smirnov, Cramer-von Mises Tests, The Annals of Mathematical Statistics, 28, 823–838, https://doi.org/10.1214/aoms/1177706788, 1957.

Davis, C., Brown, B., and Bullock, R.: Object-Based Verification of Precipitation Forecasts. Part I: Methodology and Application to Mesoscale Rain Areas, Monthly Weather Review, 134, 1772–1784, https://doi.org/10.1175/MWR3145.1, 2006a.

Davis, C., Brown, B., and Bullock, R.: Object-Based Verification of Precipitation Forecasts. Part II: Application to Convective Rain Systems, Monthly Weather Review, 134, 1785–1795, https://doi.org/10.1175/MWR3146.1, 2006b.

Davis, C. A., Brown, B. G., Bullock, R., and Halley-Gotway, J.: The Method for Object-Based Diagnostic Evaluation (MODE) Applied to Numerical Forecasts from the 2005 NSSL/SPC Spring Program, Weather and Forecasting, 24, 1252–1267, https://doi.org/10.1175/2009WAF2222241.1, 2009.

Dee, D. P., Uppala, S. M., Simmons, A. J., Berrisford, P., Poli, P., Kobayashi, S., Andrae, U., Balmaseda, M. A., Balsamo, G., Bauer, P., Bechtold, P., Beljaars, A. C. M., van de Berg, L., Bidlot, J., Bormann, N., Delsol, C., Dragani, R., Fuentes, M., Geer, A. J., Haimberger, L., Healy, S. B., Hersbach, H., Hólm, E. V., Isaksen, L., Kållberg, P., Köhler, M., Matricardi, M., McNally, A. P., Monge-Sanz, B. M., Morcrette, J.-J., Park, B.-K., Peubey, C., de Rosnay, P., Tavolato, C., Thépaut, J.-N., and Vitart, F.: The ERA-Interim Reanalysis: Configuration and Performance of the Data Assimilation System, Quarterly Journal of the Royal Meteorological Society, 137, 553–597, https://doi.org/10.1002/qj.828, 2011.

Delrieu, G., Nicol, J., Yates, E., Kirstetter, P.-E., Creutin, J.-D., Anquetin, S., Obled, C., Saulnier, G.-M., Ducrocq, V., Gaume, E., Payrastre, O., Andrieu, H., Ayral, P.-A., Bouvier, C., Neppel, L., Livet, M., Lang, M., du-Châtelet, J. P., Walpersdorf, A., and Wobrock, W.: The Catastrophic Flash-Flood Event of 8–9 September 2002 in the Gard Region, France: A First Case Study for the Cévennes–Vivarais Mediterranean Hydrometeorological Observatory, Journal of Hydrometeorology, 6, 34–52, https://doi.org/10.1175/JHM-400.1, 2005.

Descamps, L., Labadie, C., and Bazile, E.: 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, pp. 175–182, ECMWF, ECMWF, Shinfield Park, Reading, https://www.ecmwf.int/node/9015, 2011.

Descamps, L., Labadie, C., Joly, A., Bazile, E., Arbogast, P., and Cébron, P.: PEARP, the Météo-France Short-Range Ensemble Prediction System, Quarterly Journal of the Royal Meteorological Society, 141, 1671–1685, https://doi.org/10.1002/qj.2469, 2015.

Du, J., Mullen, S. L., and Sanders, F.: Short-Range Ensemble Forecasting of Quantitative Precipitation, Monthly Weather Review, 125, 2427–2459, https://doi.org/10.1175/1520-0493(1997)125<2427:SREFOQ>2.0.CO;2, 1997.

Ducrocq, V., Ricard, D., Lafore, J.-P., and Orain, F.: Storm-Scale Numerical Rainfall Prediction for Five Precipitating Events over France: On the Importance of the Initial Humidity Field, Weather and Forecasting, 17, 1236–1256, https://doi.org/10.1175/1520-0434(2002)017<1236:SSNRPF>2.0.CO;2, 2002.

Ducrocq, V., Aullo, G., and Santurette, P.: The extreme flash flood case of November 1999 over Southern France, La Météorologie, 42, 18–27, 2003.

Ducrocq, V., Nuissier, O., Ricard, D., Lebeaupin, C., and Thouvenin, T.: A Numerical Study of Three Catastrophic Precipitating Events over Southern France. II: Mesoscale Triggering and Stationarity Factors, Quarterly Journal of the Royal Meteorological Society, 134, 131–145, https://doi.org/10.1002/qj.199, 2008.

Ebert, E. E. and McBride, J. L.: Verification of Precipitation in Weather Systems: Determination of Systematic Errors, Journal of Hydrology, 239, 179–202, https://doi.org/10.1016/S0022-1694(00)00343-7, 2000.

Ehmele, F., Barthlott, C., and Corsmeier, U.: The influence of Sardinia on Corsican rainfall in the western Mediterranean Sea: A numerical sensitivity study, Atmospheric Research, 153, 451 – 464, https://doi.org/https://doi.org/10.1016/j.atmosres.2014.10.004, http://www.sciencedirect.com/science/article/pii/S0169809514003731, 2015.

Erdin, R., Frei, C., and Künsch, H. R.: Data Transformation and Uncertainty in Geostatistical Combination of Radar and Rain Gauges, Journal of Hydrometeorology, 13, 1332–1346, https://doi.org/10.1175/JHM-D-11-096.1, 2012.

Frei, C. and Schär, C.: A Precipitation Climatology of the Alps from High-Resolution Rain-Gauge Observations, International Journal of Climatology, 18, 873–900, https://doi.org/10.1002/(SICI)1097-0088(19980630)18:8<873::AID-JOC255>3.0.CO;2-9, 1998.

G. Gregoire, T., Lin, Q. F., Boudreau, J., and Nelson, R.: Regression Estimation Following the Square-Root Transformation of the Response, Forest Science, 54, 597–606, 2008.

Goovaerts, P. et al.: Geostatistics for natural resources evaluation, Oxford University Press on Demand, 1997.

Hamill, T. M.: Verification of TIGGE Multimodel and ECMWF Reforecast-Calibrated Probabilistic Precipitation Forecasts over the Contiguous United States, Monthly Weather Review, 140, 2232–2252, https://doi.org/10.1175/MWR-D-11-00220.1, 2012.

Hamill, T. M. and Whitaker, J. S.: Probabilistic Quantitative Precipitation Forecasts Based on Reforecast Analogs: Theory and Application, Monthly Weather Review, 134, 3209–3229, https://doi.org/10.1175/MWR3237.1, 2006.

Hamill, T. M., Hagedorn, R., and Whitaker, J. S.: Probabilistic Forecast Calibration Using ECMWF and GFS Ensemble Reforecasts. Part II: Precipitation, Monthly Weather Review, 136, 2620–2632, https://doi.org/10.1175/2007MWR2411.1, 2008.

Houtekamer, P. L. and Mitchell, H. L.: Data Assimilation Using an Ensemble Kalman Filter Technique, Monthly Weather Review, 126, 796–811, https://doi.org/10.1175/1520-0493(1998)126<0796:DAUAEK>2.0.CO;2, 1998.

Houtekamer, P. L., Lefaivre, L., Derome, J., Ritchie, H., and Mitchell, H. L.: A System Simulation Approach to Ensemble Prediction, Monthly Weather Review, 124, 1225–1242, https://doi.org/10.1175/1520-0493(1996)124<1225:ASSATE>2.0.CO;2, 1996.

Kai, T., Zhong-Wei, Y., and Yi, W.: A Spatial Cluster Analysis of Heavy Rains in China, Atmospheric and Oceanic Science Letters, 4, 36–40, https://doi.org/10.1080/16742834.2011.11446897, 2011.

Kain, J. S. and Fritsch, J. M.: Convective Parameterization for Mesoscale Models: The Kain-Fritsch Scheme, in: The Representation of Cumulus Convection in Numerical Models, edited by Emanuel, K. A. and Raymond, D. J., Meteorological Monographs, pp. 165–170, American Meteorological Society, Boston, MA, https://doi.org/10.1007/978-1-935704-13-3_16, 1993.

Khodayar, S., Czajka, B., Caldas-Alvarez, A., Helgert, S., Flamant, C., Di Girolamo, P., Bock, O., and Chazette, P.: 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, 2761–2780, 2018.

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

Lack, S. A., Limpert, G. L., and Fox, N. I.: An Object-Oriented Multiscale Verification Scheme, Weather and Forecasting, 25, 79–92, https://doi.org/10.1175/2009WAF2222245.1, 2010.

Lalaurette, F.: 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, 3037–3057, 2003.

Lin, Y.-L., Chiao, S., Wang, T.-A., Kaplan, M. L., and Weglarz, R. P.: Some Common Ingredients for Heavy Orographic Rainfall, Weather and Forecasting, 16, 633–660, https://doi.org/10.1175/1520-0434(2001)016<0633:SCIFHO>2.0.CO;2, 2001.

Little, M. A., Rodda, H. J. E., and McSharry, P. E.: Bayesian Objective Classification of Extreme UK Daily Rainfall for Flood Risk Applications, Hydrology and Earth System Sciences Discussions, 5, 3033–3060, https://doi.org/https://doi.org/10.5194/hessd-5-3033-2008, 2008.

Louis, J.-F.: A Parametric Model of Vertical Eddy Fluxes in the Atmosphere, Boundary-Layer Meteorology, 17, 187–202, https://doi.org/10.1007/BF00117978, 1979.

Ly, S., Charles, C., and Degré, A.: Geostatistical Interpolation of Daily Rainfall at Catchment Scale: The Use of Several Variogram Models in the Ourthe and Ambleve Catchments, Belgium, Hydrol. Earth Syst. Sci., 15, 2259–2274, https://doi.org/10.5194/hess-15-2259-2011, 2011.

Ly, S., Charles, C., and Degré, A.: Different methods for spatial interpolation of rainfall data for operational hydrology and hydrological modeling at watershed scale: a review, BASE, 2013.

Mass, C. F., Ovens, D., Westrick, K., and Colle, B. A.: Does Increasing Horizontal Resolution Produce More Skillful Forecasts?, Bulletin of the American Meteorological Society, 83, 407–430, https://doi.org/10.1175/1520-0477(2002)083<0407:DIHRPM>2.3.CO;2, 2002.

Masson, V., Le Moigne, P., Martin, E., Faroux, S., Alias, A., Alkama, R., Belamari, S., Barbu, A., Boone, A., Bouyssel, F., Brousseau, P., Brun, E., Calvet, J.-C., Carrer, D., Decharme, B., Delire, C., Donier, S., Essaouini, K., Gibelin, A.-L., Giordani, H., Habets, F., Jidane, M., Kerdraon, G., Kourzeneva, E., Lafaysse, M., Lafont, S., Lebeaupin Brossier, C., Lemonsu, A., Mahfouf, J.-F., Marguinaud, P., Mokhtari, M., Morin, S., Pigeon, G., Salgado, R., Seity, Y., Taillefer, F., Tanguy, G., Tulet, P., Vincendon, B., Vionnet, V., and Voldoire, A.: The SURFEXv7.2 land and ocean surface platform for coupled or offline simulation of Earth surface variables and fluxes, Geoscientific Model Development, 6, 929–960, https://doi.org/10.5194/gmd-6-929-2013, https://hal.archives-ouvertes.fr/hal-00968042, 2013.

Mills, G. F.: Principal Component Analysis of Precipitation and Rainfall Regionalization in Spain, Theoretical and Applied Climatology, 50, 169–183, https://doi.org/10.1007/BF00866115, 1995.

Mittermaier, M., North, R., Semple, A., and Bullock, R.: Feature-Based Diagnostic Evaluation of Global NWP Forecasts, Monthly Weather Review, 144, 3871–3893, https://doi.org/10.1175/MWR-D-15-0167.1, 2015.

Molteni, F., Buizza, R., Palmer, T. N., and Petroliagis, T.: The ECMWF Ensemble Prediction System: Methodology and Validation, Quarterly Journal of the Royal Meteorological Society, 122, 73–119, https://doi.org/10.1002/qj.49712252905, 1996.

Morin, G., Fortin, J.-P., Sochanska, W., Lardeau, J.-P., and Charbonneu, R.: Use of Principal Component Analysis to Identify Homogeneous Precipitation Stations for Optimal Interpolation, Water Resources Research, 15, 1841–1850, https://doi.org/10.1029/WR015i006p01841, 1979.

Nachamkin, J. E.: Application of the Composite Method to the Spatial Forecast Verification Methods Intercomparison Dataset, Weather and Forecasting, 24, 1390–1400, https://doi.org/10.1175/2009WAF2222225.1, 2009.

Nuissier, O., Ducrocq, V., Ricard, D., Lebeaupin, C., and Anquetin, S.: 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, 111–130, https://doi.org/10.1002/qj.200, 2008.

Nuissier, O., Joly, B., Joly, A., Ducrocq, V., and Arbogast, P.: A Statistical Downscaling to Identify the Large-Scale Circulation Patterns Associated with Heavy Precipitation Events over Southern France, Quarterly Journal of the Royal Meteorological Society, 137, 1812–1827, https://doi.org/10.1002/qj.866, 2011.

Palmer, T., Buizza, R., Doblas-Reyes, F., Jung, T., Leutbecher, M., Shutts, G., Steinheimer, M., and Weisheimer, A.: Stochastic parametrization and model uncertainty, ECMWF Technical Memorandum, p. 42, https://doi.org/10.21957/ps8gbwbdv, https://www.ecmwf.int/node/11577, 2009.

Peñarrocha, D., Estrela, M. J., and Millán, M.: Classification of Daily Rainfall Patterns in a Mediterranean Area with Extreme Intensity Levels: The Valencia Region, International Journal of Climatology, 22, 677–695, https://doi.org/10.1002/joc.747, 2002.

Pergaud, J., Masson, V., Malardel, S., and Couvreux, F.: A Parameterization of Dry Thermals and Shallow Cumuli for Mesoscale Numerical Weather Prediction, Boundary-Layer Meteorology, 132, 83, https://doi.org/10.1007/s10546-009-9388-0, 2009.

Petroliagis, T., Buizza, R., Lanzinger, A., and Palmer, T. N.: Potential Use of the ECMWF Ensemble Prediction System in Cases of Extreme Weather Events, Meteorological Applications, 4, 69–84, https://doi.org/10.1017/S1350482797000297, 1997.

Piriou, J.-M., Redelsperger, J.-L., Geleyn, J.-F., Lafore, J.-P., and Guichard, F.: An Approach for Convective Parameterization with Memory: Separating Microphysics and Transport in Grid-Scale Equations, Journal of the Atmospheric Sciences, 64, 4127–4139, https://doi.org/10.1175/2007JAS2144.1, 2007.

Ricard, D., Ducrocq, V., and Auger, L.: A Climatology of the Mesoscale Environment Associated with Heavily Precipitating Events over a Northwestern Mediterranean Area, Journal of Applied Meteorology and Climatology, 51, 468–488, https://doi.org/10.1175/JAMC-D-11-017.1, 2011.

Romero, R., Ramis, C., and Guijarro, J. A.: Daily Rainfall Patterns in the Spanish Mediterranean Area: An Objective Classification, International Journal of Climatology, 19, 95–112, https://doi.org/10.1002/(SICI)1097-0088(199901)19:1<95::AID-JOC344>3.0.CO;2-S, 1999.

Rossa, A., Nurmi, P., and Ebert, E.: Overview of Methods for the Verification of Quantitative Precipitation Forecasts, in: Precipitation: Advances in Measurement, Estimation and Prediction, edited by Michaelides, S., pp. 419–452, Springer Berlin Heidelberg, Berlin, Heidelberg, https://doi.org/10.1007/978-3-540-77655-0_16, 2008.

Schär, C., Ban, N., Fischer, E. M., Rajczak, J., Schmidli, J., Frei, C., Giorgi, F., Karl, T. R., Kendon, E. J., Tank, A. M. G. K., O'Gorman, P. A., Sillmann, J., Zhang, X., and Zwiers, F. W.: Percentile Indices for Assessing Changes in Heavy Precipitation Events, Climatic Change, 137, 201–216, https://doi.org/10.1007/s10584-016-1669-2, 2016.

Scholz, F. W. and Stephens, M. A.: K-Sample Anderson-Darling Tests, Journal of the American Statistical Association, 82, 918–924, https://doi.org/10.2307/2288805, 1987.

Schumacher, R. S. and Davis, C. A.: Ensemble-Based Forecast Uncertainty Analysis of Diverse Heavy Rainfall Events, Weather and Forecasting, 25, 1103–1122, https://doi.org/10.1175/2010WAF2222378.1, 2010.

Sénési, S., Bougeault, P., Chèze, J.-L., Cosentino, P., and Thepenier, R.-M.: The Vaison-La-Romaine Flash Flood: Mesoscale Analysis and Predictability Issues, Weather and Forecasting, 11, 417–442, https://doi.org/10.1175/1520-0434(1996)011<0417:TVLRFF>2.0.CO;2, 1996.

Shepard, D.: A Two-Dimensional Interpolation Function for Irregularly-Spaced Data, in: Proceedings of the 1968 23rd ACM National Conference, ACM '68, pp. 517–524, ACM, New York, NY, USA, https://doi.org/10.1145/800186.810616, 1968.

Stensrud, D. J., Brooks, H. E., Du, J., Tracton, M. S., and Rogers, E.: Using Ensembles for Short-Range Forecasting, Monthly Weather Review, 127, 433–446, https://doi.org/10.1175/1520-0493(1999)127<0433:UEFSRF>2.0.CO;2, 1999.

Teo, C.-K., Koh, T.-Y., Chun-Fung Lo, J., and Chandra Bhatt, B.: Principal Component Analysis of Observed and Modeled Diurnal Rainfall in the Maritime Continent, Journal of Climate, 24, 4662–4675, https://doi.org/10.1175/2011JCLI4047.1, 2011.

Toth, Z. and Kalnay, E.: Ensemble Forecasting at NMC: The Generation of Perturbations, Bulletin of the American Meteorological Society, 74, 2317–2330, https://doi.org/10.1175/1520-0477(1993)074<2317:EFANTG>2.0.CO;2, 1993.

Toth, Z. and Kalnay, E.: Ensemble Forecasting at NCEP and the Breeding Method, Monthly Weather Review, 125, 3297–3319, https://doi.org/10.1175/1520-0493(1997)125<3297:EFANAT>2.0.CO;2, 1997.

Vié, B., Nuissier, O., and Ducrocq, V.: Cloud-Resolving Ensemble Simulations of Mediterranean Heavy Precipitating Events: Uncertainty on Initial Conditions and Lateral Boundary Conditions, Monthly Weather Review, 139, 403–423, https://doi.org/10.1175/2010MWR3487.1, 2010.

Walser, A. and Schär, C.: Convection-Resolving Precipitation Forecasting and Its Predictability in Alpine River Catchments, Journal of Hydrology, 288, 57–73, https://doi.org/10.1016/j.jhydrol.2003.11.035, 2004.

Walser, A., Lüthi, D., and Schär, C.: Predictability of Precipitation in a Cloud-Resolving Model, Monthly Weather Review, 132, 560–577, https://doi.org/10.1175/1520-0493(2004)132<0560:POPIAC>2.0.CO;2, 2004.

Wernli, H., Paulat, M., Hagen, M., and Frei, C.: SAL—A Novel Quality Measure for the Verification of Quantitative Precipitation Forecasts, Monthly Weather Review, 136, 4470–4487, https://doi.org/10.1175/2008MWR2415.1, 2008.

Wernli, H., Hofmann, C., and Zimmer, M.: Spatial Forecast Verification Methods Intercomparison Project: Application of the SAL Technique, Weather and Forecasting, 24, 1472–1484, https://doi.org/10.1175/2009WAF2222271.1, 2009.

World Meteorological Organization, ed.: Guidelines on Ensemble Prediction Systems and Forecasting, 1091, WMO, 2012.

World Meteorological Organization, ed.: Guidelines on the definition and monitoring of extreme weather and climate events, Task Team on definitions of Extreme Weather and Climate Events (TT-DEWCE), 2016.